Aligning and Characterising Group Behaviours Using Role Information

by

Alina Natalia Bialkowski

B. Eng (Hons, 1st Class)

PhD Thesis

Submitted in Fulfilment

of the Requirements

for the Degree of

Doctor of Philosophy

at the

Queensland University of Technology

Image and Video Research Laboratory

Science and Engineering Faculty

2015

Abstract

With the wide deployment of visual tracking systems, a large amount of spatio-temporal data is becoming available to assist in monitoring and analysing group behaviours. However, due to the dynamic and multi-agent nature of groups, a major bottleneck restricting large-scale analysis is aligning the tracking data. The frequent role swaps between individuals within a group results in misalignment of the data and needs to be overcome before large-scale analysis can be performed.

This thesis presents research into aligning and characterising group behaviour directly from spatio-temporal data. A group can be considered as a collection of intelligent agents or autonomous entities that observe an environment and direct their activity towards achieving their goals. Before analysis can be conducted, agent positions or trajectories must be aligned. Macroscopic approaches to alignment such as density (i.e. centroids) or grid-based (i.e. occupancy maps) approaches can be used but these result in a loss of information. Microscopic approaches are preferred as they have no information loss and enable fine-grain analysis – however, continuous trajectories are generally required and finding the best template to align the data is challenging.

A major contribution in this thesis was the development of an alignment method which uses formation found directly from data using the minimum entropy data partitioning method. In addition to providing a much more compressible signal

which can be used to quickly and accurately detect group activities, it is shown that this method can be used to clean up noisy detections and can be used to provide context for tasks such as person re-identification.

The techniques and representations developed in this thesis were evaluated on sports and surveillance datasets as they provide rich sources of individual and multi-agent data for group behaviour analysis. These datasets also enable many practical applications to be demonstrated. In particular, it was shown (i) how team behaviours can be visualised and characterised through formation, (ii) how team activities can be recognised in real-time from noisy sensor data, as well as (iii) how group structure can be used to improve the accuracy of person re-identification in group situations.

Keywords

Group Behaviour, Formation, Roles, Alignment, Sports Analytics, Surveillance, Person Re-Identification, Behaviour Modelling, Occupancy Maps, Entropy, Multi Camera, Knowledge Discovery, Computer Vision, Machine Learning, Data Mining, Artificial Intelligence, Adversarial, Multi-agent.

Contents

List of Tables

List of Figures

Certification of Thesis

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher educational institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed:

Date: 27/7/15

Acknowledgments

This thesis would not have been possible without the inspiration and support of a number of people. I extend my sincere thanks and appreciation to everyone that has been a part of this journey.

Firstly, I would like to thank my supervisors Professor Sridha Sridharan, Dr Patrick Lucey, Dr Simon Denman and Associate Professor Clinton Fookes. I would not have been able to complete this thesis without their direction, ongoing feedback and advice and I am especially grateful for the time spent reviewing my articles and research over the last few years. I would like to express my gratitude to Professor Sridha Sridharan, for providing an excellent work environment at the Speech Audio Image and Video Technologies (SAIVT) lab, and for the opportunities to attend international conferences and work with great researchers. I would also like to acknowledge the financial support provided by the Queensland Government's Department of Employment, Economic Development and Innovation as part of the Smart Futures Program, and the Queensland University of Technology's Vice Chancellor's award.

During the course of my PhD, I was fortunate to have the opportunity to undertake three internships at Disney Research Pittsburgh. I would like to thank Professor Jessica Hodgins, Professor Sridha Sridharan and Dr Patrick Lucey for providing me with this opportunity as well as the admin staff and all the friends I

made there who made it such an enjoyable experience. I would like to thank the Vision Team for their insight and comments, and would like to extend a special thank you to Iain Matthews, Patrick Lucey, Peter Carr, Yaser Sheikh, and Yisong Yue for sharing their expertise and for managing to come up with new ideas and methods to evaluate every meeting. I would especially like to thank Patrick Lucey who mentored me throughout most of my PhD journey, taught me the techniques in conducting and presenting research, and continuously challenged me.

I would also like to acknowledge the past and present members of the SAIVT laboratory for the great atmosphere they created, for sharing their research expertise and for their friendship. I would particularly like to thank my colleagues in the Behaviour Analysis Group, for providing a supportive atmosphere for developing my presentation skills, and our research discussions which helped shape my work.

Finally, I would like to thank my family and friends for their support and encouragement throughout my thesis. I am eternally grateful to my parents for everything they have done for me and in helping to get me to where I am today. They will never know how much of a positive influence they have been on my life. I miss you Dad and wish you could have been here to see the completion of my PhD. To Mum, Dad, Babcia, Konstanty, Agata, Sabina, and Michael - thank you for your love and support throughout my thesis, for listening to me rehearse my work, providing me feedback, for being there through the tough times as well as providing laughter and good times to help get me through to the finish line.

Alina Natalia Bialkowski

Queensland University of Technology

July 2015

Chapter 1

Introduction

1.1 Motivation and Overview

A lot of interesting behaviours and patterns emerge when people act and move in group situations. Understanding these behaviours is important for tasks ranging from providing security and operational analytics in surveillance applications to examining strategy, individual and team performance in sports. With the wide deployment of visual surveillance and tracking systems, a deluge of visual and spatio-temporal tracking data has become available to help monitor and analyse group behaviours. Presently, such data is manually analysed by human operators which is very laborious and inherently subjective. As a result, researchers have turned to developing automated techniques to assist analysis. While advancements have been achieved in person detection, tracking and activity recognition, most of these advances have centered on individual behaviours, and analysis of the collective behaviour of groups is still quite limited.

In this thesis, a group is considered to be a collection of *agents* – autonomous

entities which observe the environment and direct their actions towards achieving their goals. While different types of groups and behaviours exist in different domains, a common way to perform analysis of their behaviours is using spatio-temporal data that represents the position and movements of each agent over time. Spatio-temporal data can be acquired from visual sources or tracking devices, and while it is easy for a human analyst to recognise patterns from such data, developing automated computer methods to represent and analyse group behaviours is challenging.

One of the reasons that has restricted the large-scale analysis of group behaviours is the difficulty in automatically acquiring continuous spatio-temporal group data. Non-invasive methods of detecting individuals such as through vision-based systems are desirable over wearable tracking devices but often result in errors such as missed and false detections and identity swaps within tracks, due to occlusions and background clutter. This makes analysis difficult because many methods of analysis rely on continuous trajectories. Such errors can be corrected over short durations, but long-term tracking is yet an unsolved computer vision problem and has restricted group behaviour analysis to short durations or to macroscopic approaches which coarsely represent a group and their global behaviours. Ideally a microscopic (fine-grained) approach which models each individual is desired as there is no information loss, however, acquiring such data is difficult.

Clean, continuous trajectories for microscopic analysis of group behaviours can be acquired by manually correcting automatically acquired data. Vision-based systems have been successfully deployed in professional sporting domains and while they still provide noisy output, the data is corrected by a team of annotators to provide continuous, clean data streams of player location information for seasons worth of matches. Despite such group behaviour data becoming more widely available, automated large-scale analysis considering individuals and the

group as a collective has been limited and a major bottleneck restricting analysis is the complexity in dealing with multi-agent data. A core analytical task involves computing the difference between groups and examples of their behaviour. This can be achieved by concatenating the features of each agent in a vector and using standard measures such as the Euclidean distance to compare vectors. However, an accurate distance measure can only be acquired with alignment of the individual agent positions or trajectories within the group setting.

In this thesis, *alignment* refers to arranging data in correct relative positions to provide feature correspondence between examples and enable accurate comparisons and analysis of the data. The dynamic nature of group movements makes this challenging to achieve, especially in long-term and large-scale analysis. For example, when comparing a group's movements at different times, any changes in their relative positions will result in misalignment as demonstrated in Figure 1.1. In this figure, despite the two examples exhibiting the exact same activity, a large distance value between the two examples is computed because agents p_1 and p_2 have swapped positions. A more accurate measure of the difference in behaviour can be acquired by aligning or re-ordering the feature vectors to match one another, because similar movements and behaviours are not defined by the identity of the agents but the positions of the agents relative to one another.

Recovering a group's structure over time is another important task for comparing groups and can be naively modelled using the distribution of each agent's position across a desired period of time as shown in Figure 1.2 (a). However, due to the dynamic nature of group movements, the relative positions of the agents changes with time and results in misalignment and overlap in the distributions. Even though a group tends to maintain a distinct spatial structure, position swaps across time make it difficult to discover this structure. The structure can be recovered by re-ordering or permuting the identities of the agents, as shown

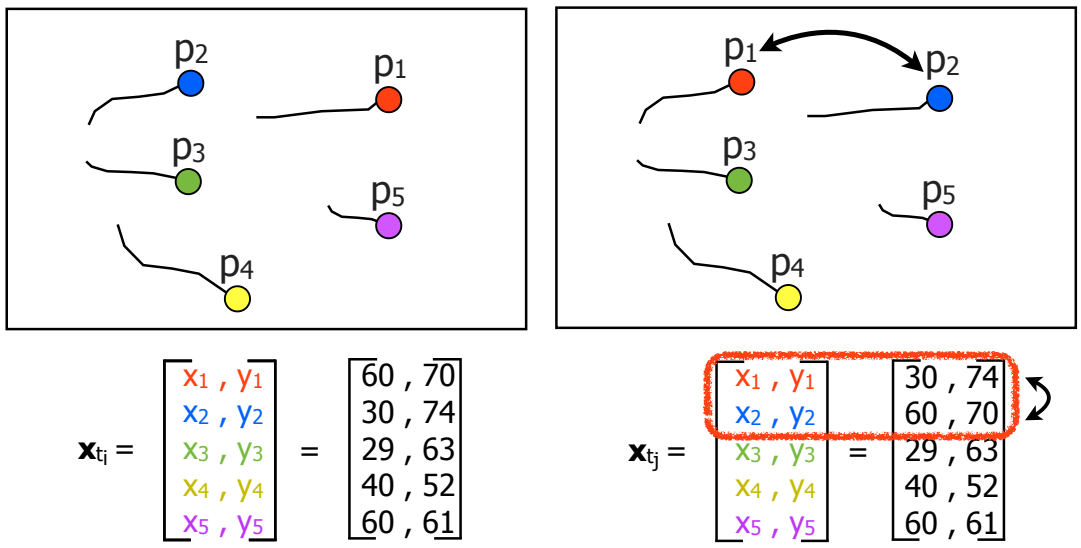
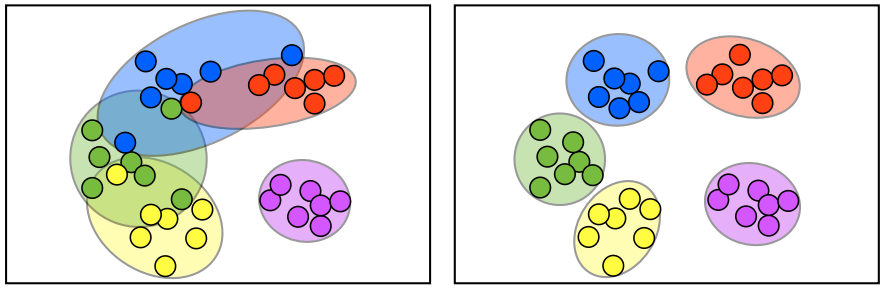


Figure 1.1: Example illustrating the importance of alignment when comparing positions or trajectories of agents across time. Between the observation on the left and right, p_1 and p_2 swap positions. Even though the group is in the same relative positions, if the original ordering of the concatenated feature vector is maintained (i.e. by agent identity, 1 to 5), a large difference is computed between the two time instants, t_i and t_j, ($||\mathbf{x}_{t_i} - \mathbf{x}_{t_j}||_2 \approx 44$ m). This can be overcome by swapping the ordering of the two vectors to match one another based on relative positions (i.e. swap p_1 and p_2), resulting in a difference of zero.

in Figure 1.2 (b). Finding alignment is challenging, as the permutations grow exponentially with the number of agents (e.g. 10 agents can be ordered in 10! ways, or $> 3.6 \times 10^6$).

Existing approaches to group behaviour analysis avoid or overcome the alignment issue in various ways. For recognising behaviours, the most common method is through the use of a dictionary of all behaviours of interest, which are used to compare observed behaviours against and sort the agents to. This allows the misalignment to be overcome but it requires prior knowledge of all possible activities and is also not suitable when evaluating long term behaviours where agents swap positions throughout the period of observation. Other approaches coarsely model the group using density and flow-based methods, particularly in crowded environments where it is difficult to detect individuals. This overcomes the alignment



(a) (b)

Figure 1.2: Example illustrating the importance of alignment when visualising group structure. (a) Looking at samples of the relative locations of the agents in a group across a period of time, there is overlap in their distributions due to relative position swaps between frames. (b) If the data is aligned at each sample through permutation (indicated by the colour of each dot above), the structure of the group can be extracted and visualised.

issue by approximating the group, but results in a loss of information as individual behaviours are not modelled. Existing approaches which consider groups at a microscopic or fine-grained level, generally avoid the alignment problem by only modelling individuals independently without taking into account the group dependency of the behaviours. This misses the important context that groups provide and the inter-dependencies in their behaviours as a collective.

A key characteristic of groups is that their movements are not random, and their behaviours are influenced by their environment and other agents around them. For example, when a group of individuals occupies a space, such as a crowd in a foyer or a gathering at a public square, recognisable patterns of interaction occur opportunistically (e.g. people moving to avoid collisions) or because of structural constraints (e.g. divergence around lamp-posts). When individuals form competitive cliques, as seen in games on a sports field, distinct and deliberate patterns of activity emerge in the form of plays, tactics, and strategies. Therefore when modelling group behaviours it is important to consider the group as a collective, the group surroundings, as well as the interaction and dependencies between agents.

1.2 Large-Scale Multi-Agent Datasets

Central to this thesis is the analysis of large-scale multi-agent datasets. Due to the advances and reduced cost of sensing technology, as well as the desire for better analysis in security, sports and commercial applications, such data is becoming more widespread.

The techniques presented in this thesis were evaluated on sports and surveillance data as these domains provide rich sources of individual and multi-agent data for group behaviour analysis. Three types of multi-agent tracking data were considered, which each provide different challenges for group behaviour analysis:

1. Continuous player tracking data and event labels from a season of professional soccer,
2. Automatically acquired player detection data from a field-hockey multi-camera system, and
3. Surveillance data from a multi-camera surveillance network.

A key insight in this thesis is that even perfect tracking data is not sufficient for understanding team behaviour, as the dynamic nature of multi-agent trajectories results in misalignment (e.g. role swaps, substitutions, and comparing different groups). For deployment in real conditions, methods which can work in real-time and on noisy data must also be developed. In surveillance data, an additional challenge is that the entire environment may not be visible at all times. For determining throughput rates of a group moving through an environment (e.g. in airports or queues), re-identifying people is important for inferring the group behaviours.

1.3 Scope of Thesis

Group behaviour analysis is a very broad research domain, and consists of tasks such as sensing and tracking agents, modelling their behaviours and interactions, as well as performing classification and prediction of activities. In this thesis, the focus is on representing and aligning multi-agent data to enable large scale analysis of group behaviours and the scope was constrained to the following objectives:

1. Representing and aligning multi-agent data to allow large-scale comparison and analysis of group behaviours.
2. Discovering a lower dimensionality subspace of groups to characterise groups and improve analysis (using clean continuous trajectories and automatically acquired noisy detection data).
3. Recognising group activities from a noisy, real-time person detection system.
4. Using group contextual information to improve analysis and better identify individuals.

The work contained in this thesis is designed to address each of these unsolved problems.

1.4 Outline of Thesis

The remainder of this thesis is organised as follows:

Chapter 2 gives an overview of the various topics related to group behaviour analysis and spatio-temporal data mining. The existing approaches to per-

forming analysis are detailed, and highlights the lack of methods to align group behaviour data.

Chapter 3 introduces a *role representation* which allows the alignment issue to be overcome, and provides a more compact representation compared to player identity. Various representations are presented and evaluated.

Chapter 4 presents a method to detect formations and roles directly from data, based on the minimum entropy data partitioning technique [83]. This provides alignment of groups and their behaviours in an unsupervised manner and enables a host of group behaviour analysis to be performed. This allows team specific characteristics to be discovered and allows team behaviour to be compared across matches throughout a whole season of data.

Chapter 5 begins the consideration of noisy data and how groups provide an important contextual cue for tasks such as cleaning up noisy detection data. In this chapter, group context is used to infer missing data by making use of the lower-dimensionality role representation, allowing the use of subspace methods such as the bilinear spatio-temporal basis model [3] to "denoise" noisy detections.

Chapter 6 presents a system to perform group behaviour analysis directly from noisy data, using a real-time detection system, and macroscopic approaches of centroids and occupancy maps.

Chapter 7 presents the utility of group information for person re-identification, which refers to re-detecting and identifying a person across different observations (e.g. due to gaps in the camera network or from occlusions). Since people often move in groups, if the group can be identified, the search space can be limited by using group context to improve performance. Group context is dependent on the domain and in team sports roles can be defined within the context of a formation and the relative positions of the players.

In a surveillance domain, different types of groups may be of interest such as families, social groups or gangs. Recognising or tracking an individual as part of a group can be more easily performed than on their own and is particularly useful when appearance features alone are insufficient for identifying individuals.

1.5 Original Contributions of Thesis

In this thesis a number of original contributions are made in the field of group behaviour representation and analysis. No other research has worked with this amount of multi-agent data before, and a major contribution was the development of an alignment procedure based on roles which enables large-scale analysis of group behaviour data. Macroscopic and microscopic approaches are proposed for aligning group behaviour data and their utility are demonstrated for analysing team behaviours in professional soccer and field-hockey analysis. When considering individuals within groups such as in surveillance, group context is an important cue and is shown to improve the important task of person re-identification, which can be used to locate individuals within group situations, correct tracking results, and facilitate group behaviour analysis. The specific contributions in this thesis are summarised as:

(i) A *role representation* to align multi-agent spatio-temporal data is proposed in Chapter 3. In the proposed role representation, the vector representing the location of each agent of a group at any time instant is re-ordered to a template, to provide a consistent representation across large datasets. This overcomes frequent role swaps which cause high variance in the data, and provides a more compressible signal for performing clustering and analysis of multi-agent data.

(ii) Three methods are proposed to align multi-agent data and are evaluated in Chapter 3 - a code book; a shape context descriptor; and normalised occupancy maps ("heat maps"). These are learnt from ground truth annotated roles, and provide a template of formation and roles from which to order agents to using the Hungarian algorithm.

(iii) In Chapter 4, an alignment procedure is proposed to learn a team's formation directly from spatio-temporal data. The method is based on minimum entropy data partitioning and reduces the variance of each role iteratively in an unsupervised manner to allow the discovery of the underlying team formation, disentangling the player distributions into distinct role distributions.

(iv) A host of new methods to characterise and compare group behaviours from large spatio-temporal datasets that only become available after aligning multi-agent data, are presented in Chapter 4:

- Discovery, visualisation and clustering of team formations
- Player analysis using group context
- Characterisation of team style from spatio-temporal data and predicting of future playing styles
- Analysis of the *home advantage* from spatio-temporal data

(v) A technique to de-noise noisy tracking data using the role representation together with a bilinear spatio-temporal basis model is developed and discussed in Chapter 5. The aligned roles are used to represent the spatial basis of the signal, and the discrete cosine transform (DCT) coefficients are used for the temporal component. This allows the underlying signal to be captured even in the presence of noise and provides a compact signal from which clustering can be performed to discover the common formations and spatio-temporal patterns of a group.

(vi) A real-time system to recognise group activities using macroscopic approaches of centroids and occupancy maps to represent and align the multi-agent data is presented in Chapter 6. These are shown to be able to detect group activities effectively even in the presence of noise.

(vii) A database for evaluating person re-identification models in real-life conditions together with an evaluation protocol to evaluate what factors affect feature performance, is presented in Chapter 7.

(viii) The use of group information to improve person re-identification using role information is proposed in Chapter 7.

1.6 Publications Resulting from Research

The following fully-referred publications have been produced as a result of the work in this thesis:

1.6.1 Book Chapters

(i) **A. Bialkowski**, P. Lucey, P. Carr, S. Sridharan, I. Matthews, "Representing team behaviours from noisy data using player role", in *Computer Vision in Sports*, T.B. Moeslund, G. Thomas, A. Hilton, Eds., Springer, Ch. 12, 2015.

(ii) S. Denman, **A. Bialkowski**, C. Fookes, and S. Sridharan, "Identifying customer behaviour and dwell time using soft biometrics", in *Video Analytics for Business Intelligence*. Springer-Verlag, 2012.

1.6.2 International Conference Publications

(i) **A. Bialkowski**, P. Lucey, P. Carr, Y. Yue, S. Sridharan, I. Matthews, "Large-scale analysis of soccer matches using spatiotemporal data", in *International Conference on Data Mining (ICDM)*, December 2014.

(ii) **A. Bialkowski**, P. Lucey, P. Carr, Y. Yue, S. Sridharan, I. Matthews, "Identifying team style in soccer using formations learned from spatiotemporal tracking data", in *International Conference on Data Mining Workshop on Spatial and Spatio-Temporal Data Mining (ICDMW-SSTDM)*, December 2014.

(iii) **A. Bialkowski**, P. Lucey, P. Carr, Y. Yue, I. Matthews, "Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviors", in *MIT Sloan Sports Analytics Conference*, March 2014. [FINALIST]

(iv) **A. Bialkowski**, P.Lucey, X. Wei, S. Sridharan, "Person re-identification using group information", in *Digital Image Computing: Techniques and Applications (DICTA)*, November 2013.

(v) **A. Bialkowski**, P. Lucey, P. Carr, S.Denman, I. Matthews, S. Sridharan, "Recognising team activities from noisy data", in *Computer Vision and Pattern Recognition Workshop on Computer Vision in Sports (CVPRW-CVSports)*, June 2013. [RUNNER UP]

(vi) P. Lucey, **A. Bialkowski**, P. Carr, S. Morgan, I. Matthews, Y. Sheikh, "Representing and Discovering Adversarial Team Behaviors Using Player Roles", in *Computer Vision and Pattern Recognition (CVPR)*, June 2013.

(vii) **A. Bialkowski**, S. Denman, P.Lucey, S. Sridharan, C. Fookes, "A database for person re-identification in multi-camera surveillance networks", in *Dig-*

ital Image Computing: Techniques and Applications (DICTA), December 2012.

(viii) S. Denman, M. Halstead, **A. Bialkowski**, C. Fookes, S. Sridharan, "Can you describe him for me? A technique for semantic person search in video", in *Digital Image Computing: Techniques and Applications (DICTA)*, December 2012.

(ix) P. Lucey, **A. Bialkowski**, P. Carr, I. Matthews, Y. Sheikh, "Characterizing multi-agent team behavior from partial team tracings: Evidence from the English Premier League", in *AAAI Conference on Artificial Intelligence*, July 2012.

(x) S. Denman, **A. Bialkowski**, C. Fookes, and S. Sridharan, "Determining operational measures from multi-camera surveillance systems using soft biometrics", in *Advanced Video and Signal Based Surveillance (AVSS)*, September 2011.

Chapter 2

Literature Review

2.1 Introduction

With the influx of data being acquired from visual sensors and tracking devices, an abundance of research has emerged to help bring understanding to data and analyse it efficiently. A lot of work has looked at analysing groups and spatio-temporal data within domains such as crowds, surveillance and sports. However, an aspect that has not been considered to a great deal is the large-scale analysis of groups at a fine-grained level, combining collective and individual behaviours to characterise and compare groups. In this chapter, relevant literature is reviewed and discussed.

2.2 Mining Spatio-Temporal Data

Mining of spatio-temporal data has received a lot of research interest in recent times due to the prevalence of location-acquisition technologies such as Global Po-

sitioning Systems (GPS), cellular networks, and Radio Frequency Identification (RFID). Using such technologies, large amounts of spatio-temporal data are being generated daily, logging movements of people, vehicles, animals and weather patterns. To efficiently analyse the volume of data being generated, a lot of research has emerged in performing efficient retrieval and automatically discovering actionable knowledge about movement behaviour for applications including transportation [131, 143], military [130], social [149], scientific studies [56] and hurricane prediction [82].

2.2.1 Trajectory Clustering

Typically, analysing large amounts of spatio-temporal data involves *clustering*, which is the unsupervised learning task of grouping objects into meaningful sets, such that objects in the same group or cluster are more similar to each other than to those in other clusters. Clustering can be used to summarise large datasets and discover dominant patterns within data, and a variety of approaches have been applied in literature to achieve this.

Compared to clustering objects and discrete data, spatio-temporal data consists of both spatial and temporal information and both dimensions must both be considered when clustering. To analyse such data, many researchers extend K-means clustering and DBSCAN ("density-based spatial clustering of applications with noise"). Many approaches simplify the task by first clustering the spatial dimension to discover a discrete set of locations from the continuous spatial coordinates, before incorporating temporal information. Ashbrook and Starner [8] used K-means clustering to find significant locations in trajectories and incorporated these into a Markov model to predict people's movements. Zhou [151] classified important places for a person from a set of their trajectories, using a

Density-and-Join-based clustering algorithm similar to DBSCAN and determined which were important using features based on the frequency of each location and the temporal distribution.

Birant and Kut [20] extended DBSCAN to consider both spatial and temporal aspects in ST-DBSCAN by incorporating additional parameters – spatial and temporal similarity thresholds for determining the neighbourhood, and a parameter to allow clusters of different densities to be discovered. Palma et al. [106] considered space and time simultaneously by clustering speed along individual trajectories using DBSCAN to determine potential locations of interest, then determined final locations based on the geography behind the trajectories. This was an extension upon the work in [6], where stops and moves were extracted from trajectories by searching for intersections between the trajectories and relevant geographic objects.

Giannotti et al. [52] extended works done in sequence mining, and provided concise descriptions of frequent spatio-temporal behaviours through the concept of "trajectory patterns", in which trajectories are defined as a set of locations with transition times between them, and proposed a method to show the cumulative behaviour of a group of moving objects. They later extended this work in [51] with a large study on mobility data mining using real-life GPS data from tens of thousands of vehicles (17000 cars during one week and 40000 cars tracked during 5 weeks). A querying and data mining system was described that facilitated the analytical process. Interesting analysis of trajectories was presented including dominant routes and patterns through two Italian cities. Chen et al. [31] also discovered popular routes from trajectories by observing the travelling behaviours of many users.

Compared to the majority of existing trajectory clustering algorithms which group similar trajectories as a whole, Lee et al. [82] proposed a method to dis-

cover common sub-trajectories which would otherwise be missed. Discovering sub-trajectories may be useful in applications where there are regions of special interest for analysis or when analysing long trajectories, and avoids having to align the data by considering only local sub-trajectories. To discover common sub-trajectories from a set of trajectories, they proposed a partition-and group framework which partitions each trajectory into a set of line segments, and forms clusters by grouping similar line segments based on density, similarly to DBSCAN. The algorithm was evaluated on hurricane data and animal tracking data, generating representative trajectories for each discovered cluster, which allowed the dominant routes to be visualised.

While DBSCAN has been very widely used for trajectory analysis, it requires the data to have well separated clusters. This is generally the case in traffic trajectories (e.g. popular locations, or cities within a map), but may not always be the case in group behaviour analysis. Morris and Trivedi [103] evaluated various trajectory distance measures and clustering techniques for determining route patterns in surveillance scenes. They found that the clustering method had little effect on the quality of results, however the performance of the distance measures was affected by the properties of the trajectories in the dataset. They found that for long trajectories, data reduction techniques worked well by focusing on coarse shape and position, but in datasets where dynamics were important, time-normalised distances performed better. The longest common sub-sequences (LCSS) distance measure performed best across most datasets as it allowed matching between trajectories of unequal lengths and was robust to noise and outliers.

2.2.2 Efficient Data Retrieval

Efficient retrieval is fundamental to performing analysis of large collections of spatio-temporal data and requires a distance measure to compare the stored data to a given query, good structural organisation of the data, and a suitable indexing method for fast retrieval.

Indexing of spatial data is a well researched field, and includes methods such as R-Trees [60], R*-Trees [16] and X-trees [19]. Spatio-temporal data indexing is more complicated, as similarity between measurements is not as well defined (i.e. spatial similarity for (x,y) data is simply the Euclidean distance in 2 dimensions, but similarity of trajectories needs to incorporate time as well). Various extensions to spatial indexing methods have been applied to incorporate temporal information. In [20], they created nodes in R-Trees for spatial objects, linked in temporal order and traversed the tree to find the spatial or temporal neighbour objects of any object. Tang et al. [131] proposed an efficient method to retrieve the k-Nearest Neighbouring Trajectories with the minimum aggregated distance to a set of query points. A "candidate generation and verification" framework was developed, using a best-first strategy and R-tree indexing was used for efficient searching. Guting et al. [59] proposed a k-nearest neighbour search on moving object trajectories, where the trajectory data was indexed in a 3D-R-tree, and a filter-and-refine strategy was employed to retrieve the data.

While data indexing is out of the scope of this thesis, it is important to note that efficiency in retrieval is achieved by pre-computing distances and sorting based on similarity of measurements. Thus, it is important to have a good data representation where similarity between different measurements is well defined.

2.3 Crowd Analysis

An area which involves analysing groups of people and their behaviours is crowd analysis. Presently, monitoring large crowds of people for suspicious behaviours and analysing crowd flow are problems that security forces and managers of large environments such as airports, train stations and sports stadiums face. These tasks are generally performed by human operators and are quite demanding, making them prone to mistakes. Automated techniques are being researched to assist in identifying coordinated activities and movements of suspicious groups of people within large crowds, as well as performing tasks such as event detection, flow estimation, and crowd simulation for improving traffic flow.

A lot of the research in crowd analysis has focussed on multi-agent tracking [4, 23, 26, 84, 107, 144]. Tracking performance for sparse crowds and medium-density crowds [4, 23, 84, 119] has achieved reasonable performance but are still poor for densely populated crowds. Compared to analysing small groups of people, in crowd analysis, the aggregate movements are often the focus of analysis. Teknomo [132] describes flow-speed-density estimation methods which aggregate pedestrian movements and compute measures of speed and density to describe the macroscopic, collective behaviours of crowds. Other approaches model the macroscopic behaviour of groups through flow models. Lin [88] computed dense flow fields to model the aggregate motion patterns of crowds and weather data from local motion observations using Lie algebra. Rather than maintaining each trajectory, flow fields are computed from the local motion observations, which makes the method robust to noisy and partially corrupted observations. Ali and Shah [4] used floor fields and a cellular automaton model for tracking in high density crowds. However, this does not consider the interactions between pedestrians. Recently, Rodriguez et al. [120] developed a data-driven approach to crowd analysis which compares "crowd patches" to a dataset which contains

similar patterns of behaviour.

Microscopic approaches involve modelling each pedestrian individually and a variety of approaches have been proposed to do this. Force-based models model individual behaviours, and are based on the assumption that the direction and speed of a pedestrian can be computed based on the combination of different forces that attract the pedestrian towards their goals but repel them from moving and static obstacles. Helbing et al. [62] proposed the social force model to model the microscopic behaviour of pedestrians. In this model, the behaviour of each pedestrian is influenced by their environment (e.g. other pedestrians and obstacles or borders) and can be represented as acceleration, attractive and repulsive forces exerted on the movement motivation of the person. Another approach is the cellular automata model, which quantises the space into a set of discrete spatial locations and dynamic potential fields are modelled. Ali and Shah [4] used this approach, however it was used for computing the aggregate behaviours.

Another microscopic approach is where each pedestrian is modelled as an "agent". Agent-based models consider each pedestrian as an autonomously acting and interacting entity. Klugl et al. [77] performed large-scale agent-based pedestrian simulation of pedestrian traffic for a railway station. The simulated pedestrians not only moved without collisions between two pre-defined locations, but were able to flexibly plan and re-plan their way through the railway station. Simulations can be useful for testing different layout options and new train schedules. Zhou et al. [150] performed collective crowd behaviours understanding by learning a mixture model of dynamic pedestrian-agents. Pellegrini et al. [107] performed tracking of pedestrians, and incorporated scene information, each pedestrian's desired destination and interactions between targets using a linear trajectory avoidance model. Tracking performance was improved in comparison to dynamic models which disregard social interaction, however the method does not model

groups of people walking together. Kitani et al. [76] modelled how people diverge around lamp-posts and other structural constraints.

Cheriyadat and Radke [32] determined dominant motions in crowded scenes by clustering partial feature trajectories. They found matching points between trajectories and clustered the trajectories using spatial and directional similarity. The tracks were iteratively clustered, beginning from the longest trajectory, creating new clusters when the distance exceeds a threshold and the results were visually evaluated.

2.4 Group Context

Recent progress in multi-agent tracking has been gained by utilising contextual features of the group [27] which can greatly reduce the solution space, making analysis and prediction tractable. Qin and Shelton [114] improved multi-target tracking via social grouping, where appearance features were used together with group information by clustering similar trajectory paths to resolve ambiguities in tracking. In other domains, contextual information has also been shown to greatly improve performance, such as in object detection [133, 147] and event detection [145]. For surveillance applications, Zheng et al. [148] made use of the fact that people generally walk in groups, and showed that representing the appearance of groups rather than individuals can be used to improve person re-identification. For analysing static groups, their structure or formation forms an important group contextual cue and can greatly reduce the search space.

2.4.1 Formations

A formation is a concept that encapsulates the structure, co-ordination and strategy of a group, and is often labelled by experts when performing analysis of team sports. Representing and detecting formations has been explored in a number of tasks including activity and pattern recognition in surveillance, recognising military tactics, and recognising team formations in sports (e.g. American football, soccer and RoboSoccer).

Tracking multiple objects moving in formation has predominantly pertained to rigid formations, such as the approach proposed by Khan and Shah [74], who classified group activities from video as having either a rigid or non-rigid formation by modelling the 3D structure of tracked participants, and determining the matrix rank required to model the structure using factorisation. Recently, Liu and Liu [91], used a mixture of Markov networks to dynamically identify and track lattice and reflection patterns in video. However, the rigid assumption falls down when considering more dynamic scenarios like tracking sports players, where the formations tend to be non-rigid (i.e. particles move freely around locally and swap positions, whilst adhering to the overall global structure).

In the sports domain, the majority of works looking into formation analysis have been in the sport of American Football. One reason for this is that American football is a lot more structured compared to continuous sports like soccer and hockey, as the game is separated into "plays" where all movement stops and the players line-up into formation. The offensive formations and movements in the sport are generally chosen from a discrete number of pre-defined set plays from the team's play-book, and recognition of formations and alignment in this domain has generally been performed using a pre-defined dictionary learnt from labelled training examples. Atmosukarto et al. [9] detected the line of scrimmage and

classified offensive team formations from broadcast footage in American Football using a multi-class linear SVM classifier using the spatial distribution of gradient intensity features extracted from video. Hess et al. [65] used a mixture-of-parts pictorial-structure model in recognising the formation of an American Football team given a top-down input image. The formations were classified by defining a set of"parts" corresponding basic player types, and a set of hard constraints based on the rules of football. The method was evaluated on 25 formation images. Pozo et al. [111] represented groups using graphical models and applied this to recognising group behaviours in European handball.

A large amount of work has looked at recognising team formations in Robot Soccer and simulated robotic soccer, to sense the opponent's strategy and direct one's robots to counteract the opponent's strategy. Various representations of a formation and data mining algorithms have been used to classify formations. Generally a formation is classified based on the (x, y) positions of the players in the team, and a training set of labelled formations. Almeida et al. [5] identified team formations in simulated robotic soccer data from game logs. All the data was labelled into one of 10 formation classes, and various supervised classification approaches were evaluated using features of the (x, y) positions of all the players (normalised by subtracting the team's centroid), and the team centroid. Reis et al. [117] and Nakashima et al. [104] both modelled formations in robo-soccer and defined a formation as a set of player positions relative to the ball. In the former, a graphical model was to define formations and plays, while in the latter, the opponent formations were learnt using Artificial Neural Networks. Ayanegui et al. [10] classified formations into one of five offensive formations using manually labelled formations to train a multi-class linear SVM. Ramos et al. [115] modelled formations in robot soccer as a planar graph, modelling the pairwise relationships between players in defensive, midfield and forward line. The 3 roles (defenders, midfielders, forwards) were learnt via k-means clustering on the x position of each

player. The planar graph allowed the structure to deform in terms of positional changes of nodes, while still preserving the topological structure and changes in formation were able to be detected when the graph was no longer planar.

While a number of works have looked at modelling formations of groups, most methods assume that there is a training set where every frame is labelled with a ground truth formation label and allows supervised learning methods to be applied. In reality it is not feasible to manually label very large datasets (e.g. season's worth of data) and it may also be desirable to discover the formations automatically, making these approaches unsuitable. Also, the observed formation may not lie in one of the existing pre-trained models, particularly in real-world data. The majority of existing methods classify formations from a single frame of (x, y) positions and do not model the variance in the player positions over time. A formation is something that is maintained over time, and hence should be modelled over a longer duration.

2.5 Sports Analysis

Sport represents a perfect test-bed for investigating group behaviour. In terms of vision research, there has been limited work in understanding group behaviour in this environment mostly due to the fact that instrumenting, capturing, processing and labelling vast amounts of video data is a costly and time-consuming endeavour. It is worth noting that an enormous amount of research interest has used broadcast sport footage for video summarisation in addition to action, activity and highlight detection [17, 43, 58, 68, 80, 92, 101, 141], but given that these approaches are not automatic (i.e. the broadcast footage is taken by a human) and that the broadcast view only captures a portion of the field, analysing group behaviour using such footage has been impossible because individuals are

normally hidden.

Sports represent a different problem to crowd analysis as it is inherently adversarial, where the microscopic movement (i.e. movement of individual agents) can give a better indication of behaviour rather than analysing macroscopic behaviour. Most current work using spatio-temporal sports data has focussed on individual behaviours thus avoiding the issue of alignment. Examples of this include work done in basketball where individual shooting, rebounding and decision-making characteristics were analysed [29, 53, 98]. Miller et al. [100] used non-negative matrix factorisation to characterise different types of shooters in basketball by modelling shot attempts as a point-process. Gudmundsson and Wolle [57] clustered the passes and movement of individual players. In soccer, Lucey et al. [94, 96], detected a team's playing style by computing an occupancy map of the team's ball movement. Pena and Touchette [108] used network theory to characterise team patterns by fixing players in their nominal position and quantifying importance based on the number of passes between players. In tennis, Wei et al. [138, 139] used Hawk-Eye data to predict the type and location of the next shot based on the behaviour of the opponent.

In multi-agent domains, the common thread of aligning trajectories has centred on using a predefined dictionary or quantised representation of the environment. The seminal work of Intille and Bobick [70] used pre-aligned trajectories to recognise a single American football play defined by a rule-based template, using a Bayesian network to model interactions between the player trajectories. Zhu et al. [152] combined the movements of the players and the ball in soccer into a single "aggregate trajectory" to classify goal scoring events into categories. Perse et al. [110] recognised activities in basketball by converting player trajectories into a string of symbols based on key player positions and actions using a quantised court. Stracuzzi et al. [126] recognised group activities in American Football us-

ing a labelled dataset of actions and matching them to the closest in the labelled dataset using dynamic time warping. Bricola [24] also used a dictionary of actions and recognised activities in basketball from player trajectories by segmented the trajectories into tracklets and matching to learnt codewords using dynamic time warping. Kim et al. [75] used motion fields to predict the future location of the ball in soccer. Carr et al. [27] estimated the centroid of team motion using real-time player detection data to predict the future location of play for automatic broadcasting purposes.

2.6 Alignment

While existing approaches in trajectory, crowd, and sports analysis successfully overcome the alignment issue by considering individual behaviours within a group, using a pre-defined dictionary, or approximating the group using density based methods, no work in the multi-agent domain has looked into aligning trajectories over long periods of time for clustering, discovering and characterising group behaviours.

In this thesis, group behaviours are modelled from traces of each agent's positions over time (i.e. trajectories). At any individual time instant, the most compact way to represent the data is to concatenate the positions of each agent into a feature vector (i.e. the (x, y) co-ordinates of each individual). However, when considering longer time durations consisting of hundreds of frames or even millions of frames of data, as is the case when analysing a season's worth of player tracking data, the dimensionality explodes and the variance is quite high, resulting in confusion between classes. With high dimensionality data, a large amount of data is necessary to train classifiers. To overcome this, feature compaction can be performed or operations to bring the data into the same space. The idea is

to maintain the same information content while minimising the feature vector dimensionality, to reduce computational cost and increase accuracy. This task can be seen as minimising the variance of the tracking data, where given the position information of multiple agents across many frames, the data is permuted to a fixed canonical template. This is similar to the idea of ensemble image alignment, where the requirement is to align all images to a canonical template.

The seminal work by Learned-Miller [81] defined the automatic alignment of an ensemble of misaligned images in an unsupervised manner as "congealing", and involves minimising the misalignment cost of a set of images to a template image by learning a parametric warp function to apply each image, such to minimise the entropy. Cox et al. [34] formulated congealing as a least-squares problem, while the RASL algorithm [109] uses rank as a measure of similarity, based on the fact that semantically similar sequences when aligned should exist within a common, low rank subspace. Other low-rank objectives, such as transformed component analysis [49] or robust parametrised component analysis [37] have also been used. More recently, methods which can deal with multiple modes (or semantically meaningful groups), have been used to simultaneously align and cluster images. The key difference between the work in image alignment compared to multi-agent data is that in this thesis, multi-agent alignment is found by computing a set of permutation matrices rather than the image warp parameters.

2.7 Summary

In this chapter, literature related to spatio-temporal data mining and group behaviour analysis were reviewed. Similarly to the motivation for this thesis, these works aim to make sense of the vast amount of data being generated in various domains and gain actionable knowledge and information from the data. The

works in these areas were highlighted as they are all connected by the fact that "aligning" or minimising the variance of the input data-streams can achieve better analysis and recognition rates. Despite this, the majority of the described approaches have avoided having to perform alignment by considering trajectories independently or coarsely modelling collections of trajectories. Other approaches have overcome the alignment issue by using a pre-defined dictionary to align the data, but must have knowledge of all possible actions and activities in advance. When modelling agents independently, the "group" aspect which incorporates interaction and dependencies between agents is lost. Approaches that aggregate local behaviours or coarsely model trajectories using density, miss out on describing the fine-grained behaviours. Compared to the described approaches, this thesis looks at analysing the fine-grained behaviours of groups by aligning the data, and incorporates the dependencies within a group using role information. In addition, this thesis considers datasets of a much greater size (i.e. 2 million frames of player tracking data), where alignment of group behaviours becomes essential.

Chapter 3

Representing and Aligning Group Behaviours

3.1 Introduction

To analyse group behaviours, a quantitative representation which characterises a group is necessary as well as a way to perform comparisons between observations. The most common approach to modelling group behaviours is using spatio-temporal data consisting of the (x,y) position of each agent across time, which can be used to describe the locations, movements and interactions of a group. Before analysis can be conducted on the data, agent positions or trajectories must be aligned. Alignment refers to providing feature correspondence between observations and reducing variance/noise so that analysis can be conducted in a common search space. In facial image ensemble alignment, the rotation, scale and translation of face images is matched to enable analysis in a common space. Similarly, spatio-temporal data must be aligned to accurately represent and compare group behaviours in retrieval, recognition and classification tasks.

In this chapter, macroscopic (coarse) and microscopic (fine) approaches to aligning spatio-temporal group data are presented and evaluated. Ideally a microscopic approach is desired where the behaviour of each individual within a group is modelled, but may not always be possible due to noisy data or difficulty in representing individuals. In microscopic approaches, group behaviour is typically represented by a vector of the positions or trajectories of each agent of the group ordered via each agent's identity. While this has been successfully applied over short durations, such an "identity" representation is problematic for large-scale analysis as it lacks common labels to compare between different data sets and across time. Even with the same agent identities, the spatial ordering within the group may differ between observations, resulting in misalignment. Macroscopic approaches, including centroids and occupancy maps which are introduced in this chapter, model groups more coarsely and can avoid the identity challenge, but result in information loss.

In this chapter a "role representation" is introduced to overcome misalignment in microscopic multi-agent data analysis, by providing common labels to perform large-scale group behaviour analysis. Because roles within a group are defined by spatial properties and relations, roles can be mathematically defined and assigned in many different ways. Three methods of representing and assigning roles are presented – codebook, shape context, and normalised occupancy map ("heat map") approaches. The three approaches are compared in role assignment experiments, with accuracy computed relative to ground truth labelled roles.

For prediction and classification tasks a smaller search space is desirable, with the aim to reduce data dimensionality while minimising information loss. In the later sections of this chapter, the macroscopic and microscopic approaches are evaluated in reconstruction and clustering experiments to compare how well they represent the underlying signal of a group and their utility in performing

large-scale group behaviour analysis.

3.2 Data for Group Behaviour Analysis

One reason that has limited large-scale group behaviour analysis at a microscopic level is the difficulty in acquiring clean tracking data. Non-invasive methods of detecting agents that don't require individuals to wear tracking devices are preferable and can be achieved with vision-based detection and tracking systems. However, automatic visual tracking is still an unsolved problem over long durations, resulting in missed and false detections, identity changes, and discontinuities in the data. This has restricted group behaviour analysis to short durations or to macroscopic approaches which observe the global behaviour of a group.

Recently, a prevalence of spatio-temporal tracking data of player and ball movement has begun to emerge in most professional sports (e.g. Prozone in soccer [113] and STATS SportsVU in basketball [125]). With manual annotation, automated tracking results are corrected to provide large amounts of clean trajectory data to enable fine-grained group behaviour analysis. In this thesis sports data is predominantly used for modelling and analysis as other sources of large-scale group behaviour data do not exist. Nonetheless, the proposed methods can be directly applied or generalised to other domains containing spatio-temporal data.

Despite the rich source of data becoming available, a major bottleneck impeding automatic group behaviour analysis is the lack of sufficient structure within the data and the complexities in dealing with multi-agent trajectory data. Although the data is organised temporally, there is no spatial ordering and a major issue centres on *aligning* individual player trajectories within a team setting.

3.3 Aligning Multi-Agent Data

Given spatio-temporal tracking data, the positions and movements of a group can be represented in a vector. Examples of group behaviour can then be compared to one another by computing the distance between corresponding points in the vector representations. If two examples are similar, the distance should be small, while dissimilar examples should result in a large distance measure. To obtain an accurate distance measure, the vectors must be of the same length and be ordered in the same way (i.e. they must be *aligned*), otherwise analysis can not be accurately performed. Macroscopic and microscopic approaches to alignment are discussed in the following sections.

3.3.1 Macroscopic Approaches

Macroscopic approaches to alignment can be used for modelling the global behaviour of a group and in situations where the identity of each agent can not be maintained or when there are missed and false detections which result in a different number of agents detected in different frames. Instead of modelling the individual agent behaviours, a macroscopic approach models the behaviours of the group more coarsely. Two macroscopic approaches examined in this thesis are centroids and occupancy maps, which are demonstrated in Figure 3.1 for two teams on a soccer field.

Centroids represent the mean position (centroid) of a group over time and can be calculated at each frame by averaging the position of the observed agents in the group (i.e. $(x_c, y_c) = \frac{1}{N}\left(\sum_{n=1}^{N} x_n, \sum_{n=1}^{N} y_n\right)$). The spread of the group can also be incorporated into the representation. Occupancy maps represent the density of agents across a fixed area, and can be calculated by splitting the environment into

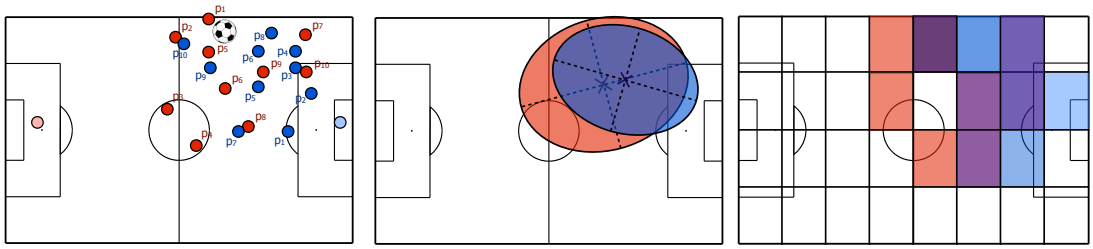
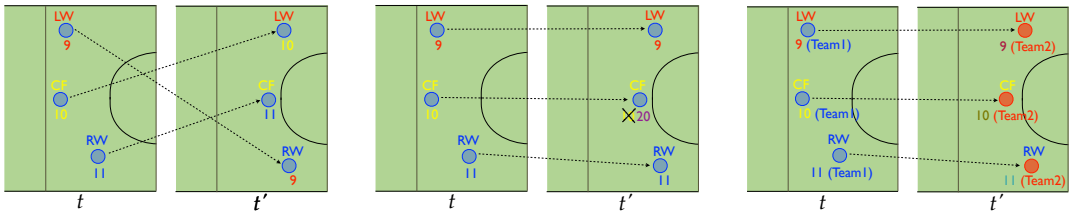


Figure 3.1: Different representations of group behaviour data. (a) The original x,y position data of each agent, (b) the centroids and spread of the two groups, (c) occupancy maps

a grid and counting how many agent detections for the group occur in each grid area. This quantises the environment into a set of discrete positions and allows the behaviour at each frame to be represented by a sorted vector of counts in each grid area, providing spatial alignment between different observations. Both approaches overcome the alignment issue without requiring the identity between observations to be maintained, but result in a loss of information. A microscopic approach which models each individual is preferable over centroids or occupancy mapping as it does not have any information loss, but requires an error-free data source.

3.3.2 Microscopic Approaches

The simplest method of aligning group behaviour from tracking data is to concatenate the (x, y) positions of each agent over time, ordered by their identity. In this identity representation, the agents are first initialised into a desired order, and remain *fixed* in this order throughout analysis. Given the continuous raw positions of N agents, their behaviour across a set of T frames can be represented



(a) Players often **switch roles**

(b) Players are often **substituted** during a match

(c) **Comparing** across teams is difficult

Figure 3.2: Challenges for representing group behaviours. The numbers represent player identities, and the labels (`LW`, `CF`, `RW`) represent the role that the player is fulfilling at that moment in time.

using a matrix of the concatenated sequence of 2D points:

$$\mathbf{D}_{T\times N} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^{(1)} & \dots & \mathbf{x}_1^{(N)} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_T^{(1)} & \dots & \mathbf{x}_T^{(N)} \end{bmatrix} \tag{3.1}$$

In this equation, $\mathbf{x}_i^{(j)} = [x_i^{(j)}, y_i^{(j)}]$ denotes the 2D coordinates of the jth agent at the ith frame, and $\mathbf{x}_i$ is the representation of all N agents for the ith frame.

To compare examples of group behaviour to one another, the distance between corresponding points in the matrix representations can be computed. However, the static ordering of agents by identity is not ideal as agents may swap positions with time, resulting in misalignment when comparing group behaviours over longer durations and across large datasets. This is particularly evident in team sports where play is dynamic and players frequently swap positions throughout the match to exploit opportunities or for other strategic reasons. Even though they may be executing the same behaviour, the identity representation will result in a large difference between two examples if players swap positions (Fig 3.2 (a)). In addition, this representation is not robust when players change due to player substitutions (Fig 3.2 (b)) or when comparing different teams (Fig 3.2 (c)).

To overcome the constant interchanging of positions and alignment issues, a *role representation* is proposed, where instead of ordering the agents be their identity, the agents are ordered based on their *role* at each frame defined by their position relative to the other roles.

In team sports, there is already a well established vocabulary for naming roles that incorporate both spatial and strategic aspects (e.g. in soccer the left-wing plays in-front of the left-back and to the left of the centre-midfielder). The roles can be encapsulated in a *formation* which is a spatial arrangement of players and can be defined as the set of roles. A formation is effectively a strategic concept and different teams can use the same formation simultaneously. Given a defined formation or set of roles, the agents can be assigned a role at each frame to overcome the issue with misalignment and provide a consistent ordering to enable large-scale analysis of behaviours.

3.4 Role Assignment

The goal is to infer which role each player is fulfilling at each time instant, from a set of roles defined within a formation, $\mathcal{F}$.

Definition 3.4.1 *A* formation $\mathcal{F}$ *is an arbitrarily ordered set of* N *roles* $\{R_1,\ R_2, \ldots,\ R_N\}$ *which describes the spatial arrangement of* N *players.*

Each role within a formation is unique (i.e. no two players in a team can have the same role at the same time), but players can swap roles throughout the match. Additionally, multiple formations may exist which consist of different sets of roles. Essentially, the assignment of N roles to a set of N players can be interpreted as applying a permutation matrix to the identity representation at each frame, such

that the role-representation at each frame is given by:

$$\mathbf{r}_t = \mathbf{P}_t \mathbf{x}_t. \tag{3.2}$$

At each time instant, the positions of the players are permuted so that the role ordering is maintained. This is demonstrated in Figure 3.3.

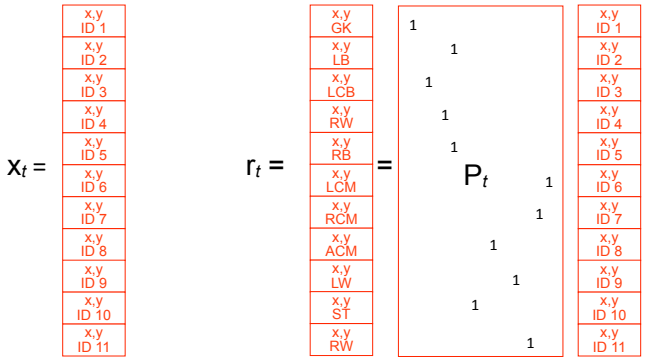


Figure 3.3: Role assignment can be seen as applying a permutation matrix to each frame of the original data ordered by identity.

The permutation matrix $\mathbf{P}_t$ can be found by minimising the total cost of the player positions to a template formation. This is a combinatorial optimisation problem, or more specifically a "linear assignment problem" between identities and roles which can be efficiently solved in polynomial time using the Hungarian algorithm [79].

Because roles are defined by spatial properties and relations, the template formation can be defined in many ways, and it is possible to infer roles for a set of (x, y) locations in multiple ways. In this chapter, supervised approaches are proposed in which the formation is learnt from a dataset of ground truth labelled roles. The overall role assignment procedure is shown in Figure 3.4, indicating three types of descriptors which can be used to determine the optimal assignment of roles to player positions, and these are described in the following sections.

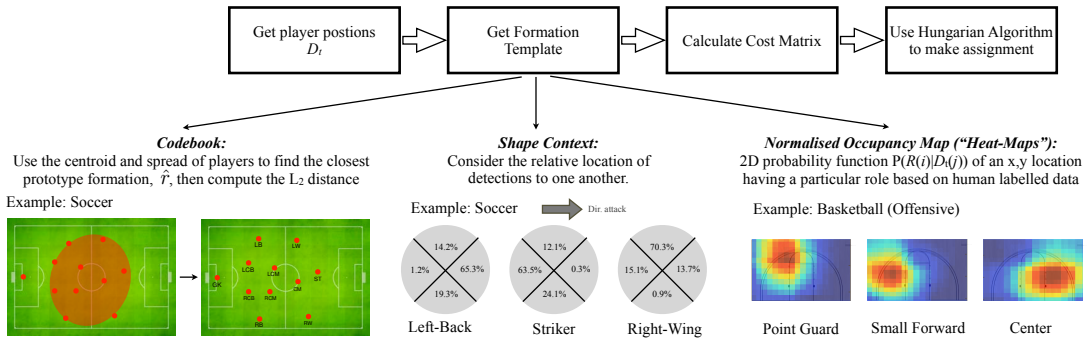


Figure 3.4: Role assignment procedure

3.4.1 Codebook

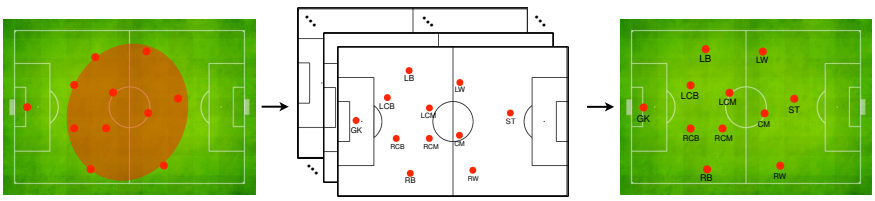


Figure 3.5: Codebook role assignment

If the formation is known, the simplest method of describing a set of roles is through a codebook or set of labelled exemplars from the training set, $\mathcal{F} = [\mathbf{r}_1, \ldots, \mathbf{r}_M]^T$, where $\mathbf{r}$ is an exemplar of that formation and M is the number of exemplars. The template exemplar, $\hat{\mathbf{r}}$, can be selected by finding the most similar exemplar to the input detections. As the input detections are in an arbitrary order, they can not be directly compared to the labelled exemplars without testing each permutation. Instead, the mean and range of the input detections are used to compare against the exemplars. The template is then set as the example in the codebook with the minimum distance from the input detections, or by training a regressor to predict a likely exemplar [95].

Given the closest exemplar in the codebook, $\hat{\mathbf{r}}$, the cost matrix can be determined by computing the Euclidean distance between each player position and

each prototype role position as follows,

$$\mathrm{C}^{\mathrm{CB}}(i,j) = \|\mathbf{x}_t(i) - \hat{\mathbf{r}}(j)\|_2 \tag{3.3}$$

where i and j refer to the ith player identity and jth role in the exemplar formation.

3.4.2 Shape Context

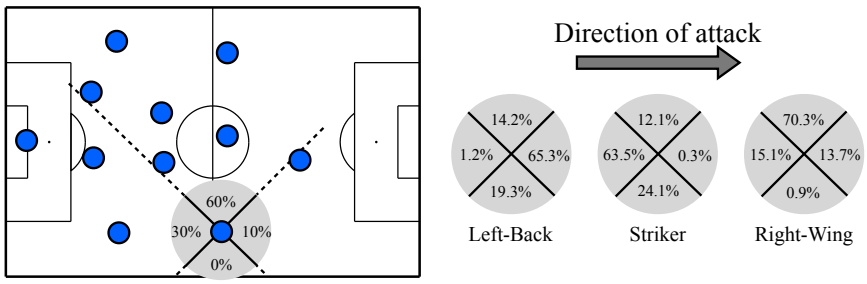


Figure 3.6: Shape context role assignment. (a) A shape context descriptor for a given player can be calculated as the percentage of players located in each angle band relative to the player. (b) Example shape context descriptor templates are shown above for three roles. These are learnt by computing the mean shape context descriptor for the given role across the training data.

The idea of shape context [18] was adapted to evaluate the likelihood of assigning a particular role to a given (x, y) player location. The key idea here is that instead of using absolute position, the locations of all other players $\{\mathbf{x}_t \setminus \mathbf{x}_t(i)\}$ in coordinates *relative* to $\mathbf{x}_t(i)$ are considered. The motivation is that roles are defined by relative location in terms of angle. For example in soccer, the right-wing should always be in front of and to the right of all other players. Equivalently, all players should be behind and to the left of the right-wing.

A descriptor $G\big(\mathbf{x}_t(i)\big)$ is computed for each player $(x, y)_i$ in $\mathbf{x}_t$ by computing the locations of the remaining players in $\mathbf{x}_t$ relative to $(x, y)_i$, as in Figure 3.6 (a).

The relative displacements are then quantised in terms of angle. As a result, four bins are formed: in front, behind, left and right. An exemplar descriptor $G\big(\mathcal{R}(j)\big)$ for each role is learnt by computing the mean of the labelled training data descriptors, with examples shown in Figure 3.6 (b). The cost of assigning a particular role to a detection is based on the distance between the descriptor generated for the $(x, y)_j$ location and the learned exemplar descriptor for the hypothesised role,

$$\mathrm{C}^{\mathrm{SC}}(i, j) = d\Big(G\big(\mathbf{x}_t(i)\big), G\big(\mathcal{R}(j)\big)\Big) \tag{3.4}$$

where the chi-squared distance measure was used as the distance function.

3.4.3 Normalised Occupancy Maps

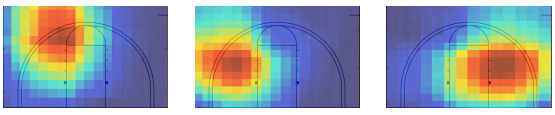


Figure 3.7: Normalised occupancy maps ("heat maps") provide a probabilistic distribution of each role's location for performing role assignment. Example heat maps for three basketball roles are shown above.

Alternatively, a probability distribution for each role within the formation can be used. These are normalised occupancy maps or "heat-maps" because when visualised, they appear to be hot in areas of the field/court where the player is most likely to be as shown in Figure 3.7. The formation can be described as a set of probability distributions, $\mathcal{F} = [\mathbf{p}_1, \ldots, \mathbf{p}_N]^T$, where $\mathbf{p}_n$ is the vectorised 2D probability function of the nth role. The 2D probability distribution function of each role is learnt by dividing the playing surface into a collection of discrete cells and generating frequency counts of how often each role occupies each cell based

on human labelled exemplar data. Given a set of detections, the probability of each individual role assignment to each (x, y) location in $\mathbf{x}_t$ is calculated as:

$$\mathrm{C}^{\mathrm{HM}}(i,j) = -\log \mathrm{P}\Big(\mathcal{R}(j)\big|\mathbf{x}_t(i)\Big) \tag{3.5}$$

and the results are combined to generate an energy reflecting the probability of the assignment $\mathbf{r}_t$ of roles.

3.4.4 Role Assignment Accuracy

The three role representations were compared in terms of role assignment accuracy using player tracking data from two sports datasets. Both datasets were taken from men's professional leagues across a season: one was from basketball, and the other was from soccer, and an inventory of the data is given in Table 3.1. To validate the different role assignment approaches, a random number of single frames were labelled for formation roles by an expert. As role depends on the team formation, the datasets for each sport were selected so the formations were constant: a 2-3 zonal formation in basketball, and a 4-2-3-1 formation in soccer. For the various descriptors, the annotated frames were broken into two partitions for cross-validation.

The results for each formation descriptor are presented in Table 3.2. It can be seen that the heat-map outperforms the other descriptors. To overcome the small amount of annotated frames[1], the heat-maps were blurred with a Gaussian filter. With more examples, it is expected that the codebook would achieve similar performance. The poor performance of the shape-context descriptor can be attributed to it not being robust to non-rigid deformations. A performance of

[1] it takes 30 seconds to annotate a frame for basketball and 1 minute for a frame of soccer

Dataset	Games	Total Frames	Total Time	Frames Role Annotated
Basketball	600 (2400 quarters)	47,790,000	28,800 mins (480 hours)	534
Soccer	354 (708 halves)	17,280,000	31,860 mins (531 hours)	403

Table 3.1: Inventory of the data used for basketball and soccer.

	Accuracy of Approach		
Dataset	**Codebook**	**Shape-Context**	**Heat-Maps**
Basketball	65.62	74.10	**89.71**
Soccer	73.07	57.84	**89.55**

Table 3.2: Accuracy of role assignment using the three descriptors. The results were obtained on the frames manually annotated for role (534 for basketball and 403 for soccer).

approximately 90% on both datasets using the heat-maps is close to the upper limit as the reliability between different annotators would likely be the same due to ambiguity in roles.

3.5 Reconstruction Experiments

For prediction and classification tasks, a smaller search space is desirable, with the aim to reduce data dimensionality while minimising information loss. This is particularly important for representing temporal information, where the dimensionality explodes and can make analysis infeasible. In this section, the macroscopic and microscopic approaches to aligning group behaviour data were evaluated in terms of their compressibility using tracking data from a season of professional soccer from Prozone [113]. Ten second clips of every shot on goal from the season (excluding those where players had been sent off) were taken as the dataset,

resulting in 9580 examples of group behaviour.

A common method of dimensionality reduction is Principal Component Analysis (PCA), in which basis are learnt to represent the data, and the original data is represented as a linear combination of the basis. The basis are ordered by variance, so that those which represent more of the variation in the data are sorted first. By projecting down into the bases representing the majority of the variance in the signal, the dimensionality of the signal can be reduced while maintaining the majority of the information content. This also gives an indication of the redundancy in the signal.

Because the spatial relationships of a formation are defined in terms of roles and not by individualistic attributes like the identity of players (who frequently swap roles during the game), it is expected that the spatio-temporal patterns in the role representation $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_T\}$ will be more compact compared to the identity representation $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_T\}$. Three different permutations of the data are considered. Ordering by:

1. *Identity* – order the players to a template at the start of the match and remain in these static roles throughout
2. *Roles(1)* – order the players to a formation template at every frame
3. *Roles(2)* – order the players at the start of each shot on goal snippet

The different representations were evaluated based on reconstruction error, i.e. how well the low dimensional representations matched the original data using the L_2 norm of the residual, $\Delta\mathbf{r} = \hat{\mathbf{r}}_t - \mathbf{r}_t$. The PCA reconstruction results are presented in Figure 3.8. It can be seen that ordering detections by role gives a much lower reconstruction error compared to the identity representation, indicating that this representation provides better alignment between examples

and allows better basis to be learnt for representing the underlying spatial signal. When reconstructing trajectories (i.e. the bottom row of the plots), the *Roles(2)* approach performs better compared to re-ordering the players at every frame in *Roles(1)* ordering, as temporal continuity is maintained (i.e. there is temporal redundancy which allows the use of less principal components to represent the signal).

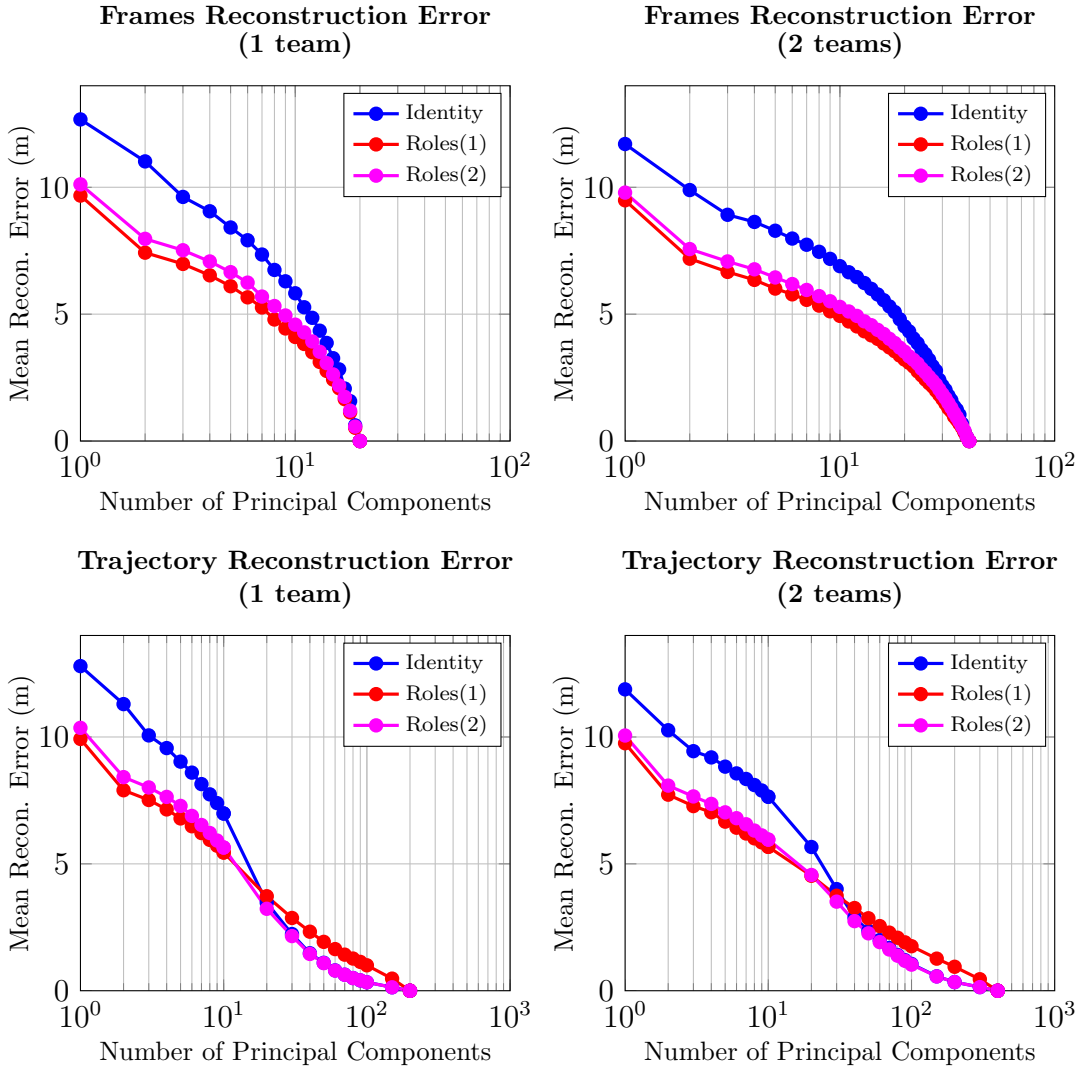


Figure 3.8: PCA reconstruction of frames and trajectories for one and two teams.

Next, the macroscopic approaches approaches were evaluated. Using the same data, the centroid and spread for each frame was extracted, forming the *centroid*

macroscopic representation of the data. This data was used to evaluate how well the original signal can be reconstructed (i.e. reconstructing the original x,y positions of the players, given just the centroid and spread of the team). This was performed using linear regression, using half the data for training and the other half for testing. In addition to reconstructing the original signal, the mean-removed signal was also reconstructed to evaluate how much of an influence the centroid or mean of the team's formation plays in the signal. The results are presented in Table 3.3.

Representation	**Reconstruction Error (m)**			
	Frames		Trajectories	
	1 Team	2 Teams	1 Team	2 Teams
Identity	10.68	9.12	10.38	8.95
Identity (mean-removed)	10.97	9.32	10.76	9.20
Roles(1)	**6.89**	**6.23**	**6.76**	**6.11**
Roles(1) (mean-removed)	7.18	6.43	7.06	6.34
Roles(2)	7.49	6.65	7.37	6.53
Roles(2) (mean-removed)	7.77	6.88	7.68	6.80

Table 3.3: Reconstruction error when using linear regression to reconstruct the (x,y) positions from centroid and spread. The best performance in each column is highlighted in bold.

From Table 3.3, it can be seen that the mean-removed reconstruction performed worse than the non-mean removed data. This shows that the mean (i.e. the team centroid, which represents where the team is located on the field) provides context to better represent team structure. Also, the reconstruction of data for 2 teams performs better than for 1 team, as there is correlation between the positions of the two teams, which provides additional context. The *Roles(1)* role representation, where the frames are re-aligned at each frame, performed best in reconstructing the data, and indicates that this alignment method best represents the underlying team structure. A similar error value was achieved when using 4 principal components in the previous experiment, which shows that while centroids are a simple representation, they represent a significant proportion of

the signal.

Next, the occupancy map representation was evaluated. In Figure 3.9, the quantisation error was evaluated on the dataset for various descriptor sizes (coarse to fine), which defines how many field areas or 'bins' to split the field into. It can be seen that a large dimensionality is required to reduce the quantisation error (e.g. over 500 bins are required to limit the quantisation error to 1.5 m). The compressibility and reconstruction error of the Occupancy Map representation was then evaluated using PCA using a 30×18 descriptor, consisting of 540 dimensions per team. This was selected as it only has a quantisation error of 1.4 m per player which gives reasonable precision for performing analysis. When reducing the dimensionality with PCA, a significant blurring of the occupancy maps results. Rather than computing the L_2 reconstruction error, a modified L_1 reconstruction error was used to give a more intuitive representation of the error, in the presence of such blurring. The measure counts how many players differ between the two representations, relative to the original signal, differing from the regular L_1 distance in that the error is only computed for bins which contained a value in the original representation. This results in 0 error for a perfect reconstruction, and a maximum error of 10 (i.e. all 10 players). The reconstruction results are presented in Figure 3.10. It can be seen that the dimensionality of the signal is extremely large, especially when reconstructing tracks (i.e. a concatenation of 10 frames = 5400 dimensions per team). It is evident that while this approach allows for alignment without requiring identity of the agents or players, it requires a large dimensionality to accurately represent behaviours.

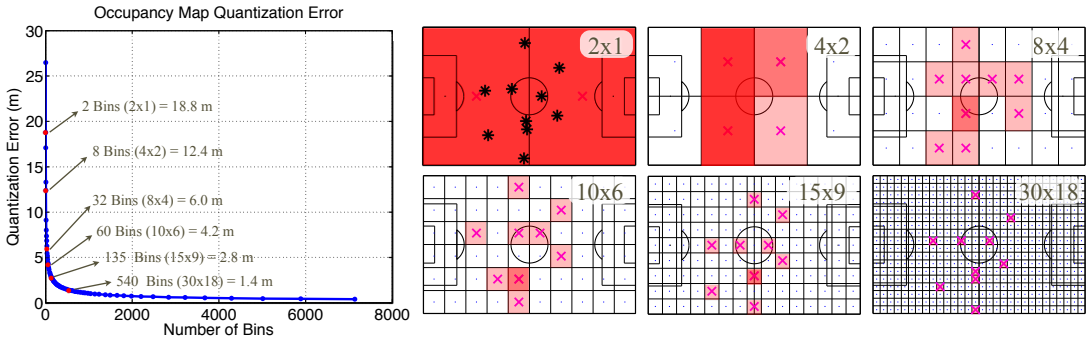


Figure 3.9: Quantisation error of the occupancy map representation

3.6 Clustering Experiments

Next, the different representations were evaluated in clustering experiments to see how a small dictionary of examples could represent the full dataset. This is a common form of analysis that is performed on large datasets to discover dominant patterns and meaningful groups or sets of data. A good representation should have a low within-clustering distance. K-Medoids clustering was used on the set of 9580 10-second shot snippets to get a set of exemplar trajectories that best represent the types of shots that teams execute. The average Euclidean distance between corresponding points on the two trajectories was used as the distance measure.

The occupancy maps were clustered using the Earth Mover's Distance (EMD) [121]. This means that instead of the distance measure counting how many occupancy spaces are changed, a measure of how far they were displaced is given. This provides a more comparable measure to the other representations, but is computationally expensive. The EMD computes the distance between two probability densities, and can be calculated between two normalised histograms **a** and **b** as the solution of the transportation problem:

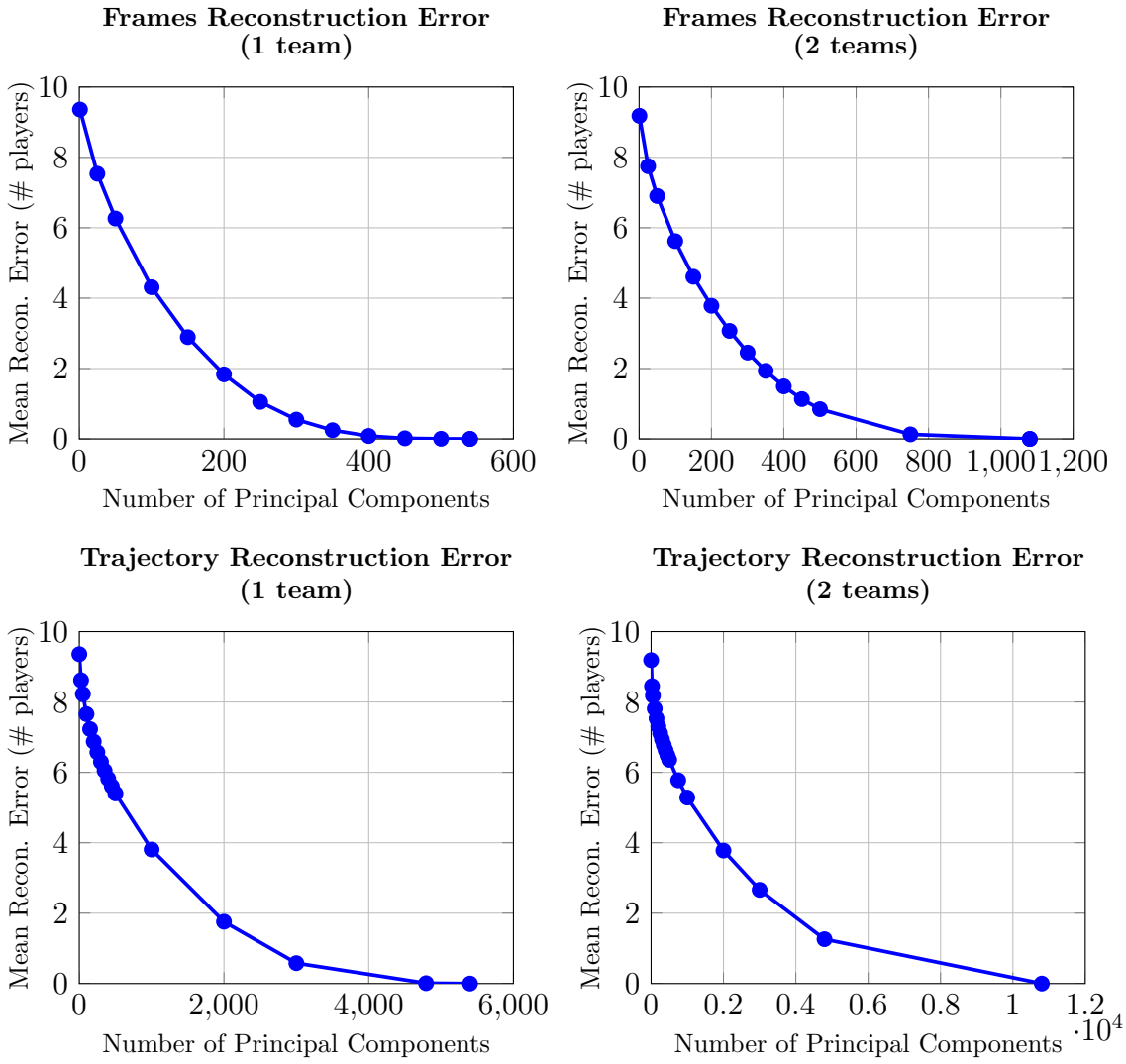


Figure 3.10: PCA reconstruction using the Occupancy Map representation

$$\min_{f_{qt}\geq 0} \sum_{q,t=1}^{D} d_{qt} f_{qt} \quad \text{s.t.} \sum_{q=1}^{D} f_{qt} = a^t, \sum_{t=1}^{D} f_{qt} = b^q. \tag{3.6}$$

where the variable f_{qt} denotes a flow representing the amount transported from the qth supply to the tth demand (i.e. how many players are transported) and d_{qt} the ground distance (i.e. how far the players are displaced).

Clustering was performed in each representation's own space, but the results were

evaluated in a common space using the original (x,y) data ordered by roles at the start of each snippet (i.e. the *Roles(2)* representation, which was found to be best for representing trajectories in Fig 3.8). The reconstruction error was calculated as the average distance between each example and its cluster centre, and reported per detection. The clustering results are presented in Figure 3.11.

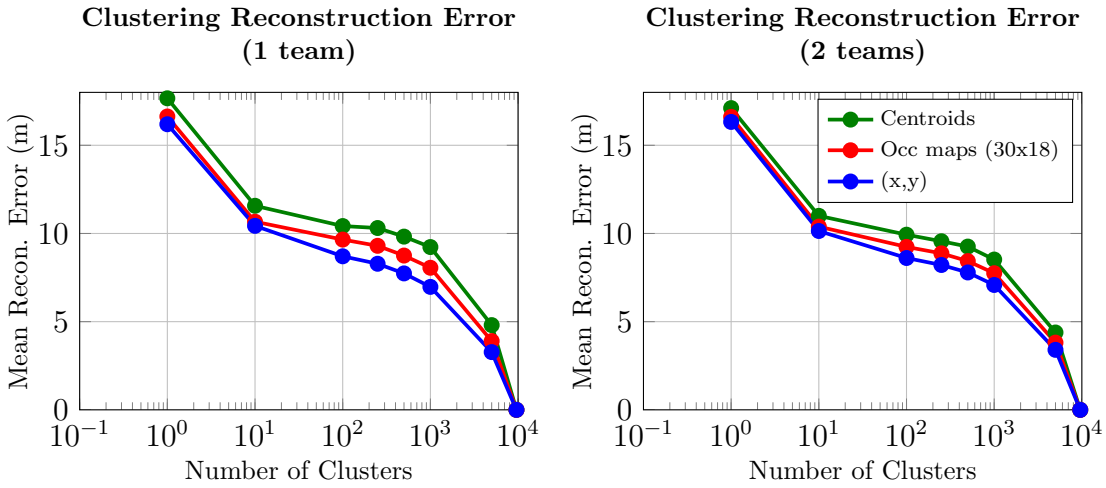


Figure 3.11: K-medoids clustering results using different representations.

From the clustering results, it can be seen that there is quite high variability in shots and a high number of examples are necessary for the cluster centres to well represent the examples assigned to them (i.e. low mean reconstruction error in the plot). An initial knee-point is visible at around 10 clusters, indicating that there appear to only be a few coarse types of shots. These could correspond to commonly known shot classes taken from different locations such as: open-play; a counter-attack (where players break quickly from one-end to the other); corners; penalties; and free-kicks. Despite this, the error is still quite high, which shows that there is great variability in how the players are arranged and move. Comparing the different representations, the original (x,y) data ordered by roles performs best followed by occupancy maps, with centroids performing the poorest. This is expected as the centroids have the greatest information loss (and hence poorer clustering results). Thus, a microscopic approach using roles is ideal for analysis.

3.7 Summary

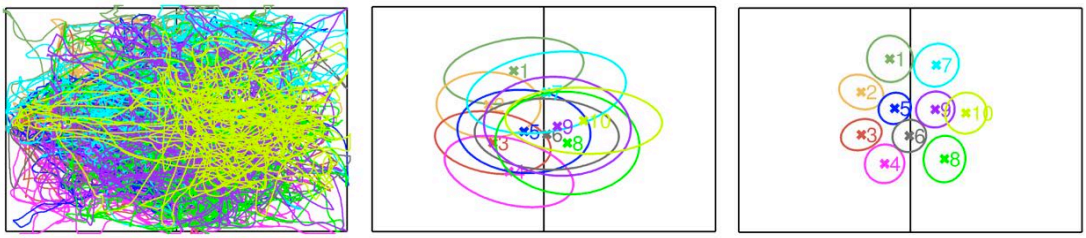
In this chapter, various methods of aligning group behaviour were presented and evaluated. Macroscopic approaches of centroids and occupancy maps were shown to be able to represent group behaviours, but at the cost of information loss. Despite this, centroids were shown to still represent a large portion of the underlying signal and can be used to provide context. Occupancy maps were shown to be able to represent behaviours on a coarse or fine basis, but at the cost of dimensionality. Ideally a microscopic approach is desired, as there is no information loss. It was shown that the static assumption of ordering players by identity is not ideal as there is constant interchanging of positions making the dimensionality of the resulting subspace much higher. To overcome this, a role representation was proposed and three methods of assigning roles were presented, with the heat map approach performing best. The role representation was shown to best represent the data in reconstruction and clustering experiments, demonstrating its ability in enabling large-scale analysis of multi-agent behaviours.

Chapter 4

Characterising and Visualising Group Behaviours

4.1 Introduction

Despite a large amount of player and ball tracking data becoming available in professional team sports, large-scale mining of such data has been limited due to the difficulty in representing dynamic multi-agent trajectories. A major issue centres on aligning player positions over time, and is apparent when looking at the distribution of player positions across a match. In Figure 4.1 (a) and (b) the trajectories and distributions of soccer players across a 45 min match half are shown, demonstrating how the continuous interchanging of player positions results in significant overlap in the player distributions. In Chapter 3, a *role-based* approach to alignment that dynamically updates each agent's role at each frame was presented to overcome the misalignment. However, this required prior knowledge of the structure or formation that the group adopts, consisting of a set of pre-defined roles. There may not exist a single template that all groups



(a) (b) (c)

Figure 4.1: Player swaps throughout a match cause misalignment in the data that need to be overcome to perform large scale-analysis. (a) Given the trajectory of each player during a match half, it can be seen that players continually swap positions. (b) The distributions of the player positions over the half highlights the overlap. (c) Using the proposed iterative role-assignment procedure, a role label is assigned to each player at the frame-level allowing the underlying structure or *formation* of the team to be extracted and visualised.

adopt and it may not be possible to acquire this in advance. For example, in soccer there are many formations that teams can utilise such as a 4-4-2, (i.e., four full-backs, four mid-fielders and two strikers) or a 4-2-3-1, (i.e., four full-backs, two holding midfielders, three attacking mid-fielders and a striker). Each formation has a different structure in which different positional responsibilities or roles are assigned and these are important to distinguish when comparing teams and strategies. Therefore, a single template cannot be used to align player positions across all teams, particularly when attempting to discover the specific style employed by each team within each match.

In this chapter, a method is proposed to learn a group's formation directly from data based on the minimum entropy data partitioning method [83, 118]. This disentangles tracking data into distinct role distributions (such as in Fig. 4.1(c)), allowing a group's underlying formation to be discovered as well as providing alignment for improved large-scale analysis of group behaviours. Using this approach, the unique characteristics of a sports team can be visualised in terms of the formation descriptor which encapsulates the team's structure, position and movements. This provides a strong cue for team identity and can be used to

find teams which play similar styles, or find the different styles a team adopts in different circumstances (e.g. a team may play one style at home and another away, or one style against a top-team and another against a bottom team). Prior knowledge of a team's playing style is also very useful for predicting future behaviours and planning strategies against future opponents. In this chapter, the procedure to discover formations from tracking data is presented and the utility of the approach is demonstrated for group behaviour analysis tasks using a full season of player and ball tracking data from a professional soccer league ($>$ 400 million data points).

4.2 Data: Player Tracking in Soccer

For this work, an entire season of soccer player tracking data from Prozone [113] was utilised. The data consists of 20 teams who played home and away, totalling 38 matches for each team or 380 matches overall. Six of these matches were omitted due to missing data. The 20 teams are referred to using arbitrary labels $\{A,\ B,\ \ldots,\ T\}$. Each match consists of two halves, with each half containing the (x, y) position of every player at 10 frames-per-second. This results in over 1 million data-points per match, in addition to the 43 possible annotated match events (e.g. passes, shots, crosses, tackles etc.). Each of these events contains the time-stamp as well as location and players involved. An inventory of the tracking data is given in Table 4.1, and a list of events annotated in each match is given in Table 4.2.

Statistic	Frequency
Teams	20
Matches	374
Frames	21.5M
Data Points	480M
Ball Events	981K

Table 4.1: Inventory of the soccer dataset used for this work.

Pass	Foul - Direct FK	Cross	Catch Drop Save
Pass Assist	Foul - Indirect FK	Cross Assist	Catch Save
Corners	Foul - Penalty	Reception	Punch
Shot on Target	Foul - Throw-in	Reception Assist	Punch Save
Shot off Target	Offside	Reception Save	Diving
Goal	Yellow Card	Catch	Diving Save
Own Goal	Red Card	Catch Drop	Drop of Ball
Neutral Clear Save	Running with Ball	Chance	Substitution
Block	Drop Kick	Pass Save	Hold of Ball
Clearance Uncontrolled	Neutral Clearance	Player Out	Clearance

Table 4.2: List of match statistics used to describe team behaviour.

4.3 Discovering Formations from Data

In team sports like soccer, there is an inherent global structure that the players adhere to termed a *formation*. This is effectively a strategic concept which defines how the team distributes its players across the field in an aim to maximise their chances of winning while trying to minimise the chances of their opposition. Even though players can actively change roles during a match (altering who occupies each role on a per frame basis), they tend to adhere to a formation. The long-term spatial structure of a team, or group in general, could intuitively be represented by the mean location of its agents. However, with constant swapping of positions, the mean positions of the agents won't represent the true spatial structure employed by the group. The role swapping must be overcome to discover the formation.

In this section, the task of learning a model of a group's formation directly from tracking data is discussed.

Mathematically, a *formation*, $\mathcal{F}$, can be defined as an arbitrarily ordered set of N *roles*, $\{R_1,\ R_2, \ldots,\ R_N\}$, which describe the spatial arrangement of N agents of a group. While roles can be represented in a number of ways, in Chapter 3 it was found that a "heat-map" approach in which each role is represented by a probability density function of location performs best. Using this representation, the best estimate of the underlying formation of a group from tracking data $\mathbf{D}$ is equivalent to finding the most probable set $\mathcal{F}^*$ of 2D probability density functions,

$$\mathcal{F}^* = \arg\max_{\mathcal{F}} P(\mathcal{F}|\mathbf{D}). \tag{4.1}$$

Criteria must be defined for which formation (i.e. set of role probability density functions) is more likely for a given set of tracking data in order to solve this equation. To begin, the 2D probability density function $P(\mathbf{X} = \mathbf{x})$ which models the tracking data $\mathbf{D}$ is considered. In other words, $P(\mathbf{x})$ represents the heat-map for the entire group (or team) which can be modelled as a linear combination of the heat maps for each role,

$$\begin{aligned} P(\mathbf{x}) &= \sum_{n=1}^{N} P(\mathbf{x}|n)P(n) \\ &= \frac{1}{N}\sum_{n=1}^{N} P_n(\mathbf{x}). \end{aligned} \tag{4.2}$$

To discover the spatial structure of a group, each role should be distinct and well separated from other roles. This is analogous to team sports where players need to strategically spread out so that the entire field is adequately covered and so that different players are responsible for different portions of the field. To achieve distinct roles, the probability density functions of each role within a group should exhibit minimal overlap with one another. This is equivalent to minimising the

overlap of each role probability density function with the group's probability density function. Following the ideas of minimum entropy data partitioning [83, 118], Kullback-Liebler divergence was employed to measure the overlap between two probability functions $P(x)$ and $Q(x)$,

$$KL\left(P(x)\|Q(x)\right) = \int P(x)\log\Big(\frac{P(x)}{Q(x)}\Big)dx. \tag{4.3}$$

Since divergence is a strictly positive quantity (and completely overlapping probability density functions have zero divergence), a penalty V_n is employed based on the negative divergence value between the heat map $P_n(\mathbf{x})$ of an individual role and that of the whole group $P(\mathbf{x})$,

$$V_n = -KL\big(P_n(\mathbf{x})\|P(\mathbf{x})\big). \tag{4.4}$$

Computing the optimal formation $\mathcal{F}^*$ is equivalent to determining the optimal set $\mathcal{F}^* = \{P_1(\mathbf{x}),\ldots,P_N(\mathbf{x})\}^*$ of per-role probability density functions $P_n(\mathbf{x})$ that minimise the total overlap,

$$\mathcal{F}^* = \arg\min_{\mathcal{F}} V. \tag{4.5}$$

Substituting the expressions for KL divergence into the total overlap cost illustrates the dependence on each role-specific 2D probability density function

$$\begin{aligned} V &= \sum_{n=1}^{N} P(n)\Big(-KL\big(P_n(\mathbf{x})\|P(\mathbf{x})\big)\Big) && (4.6)\\ &= -\sum_{n=1}^{N} P(n)\int P_n(\mathbf{x})\log\Big(\frac{P_n(\mathbf{x})}{P(\mathbf{x})}\Big)dx && (4.7)\\ &= -\sum_{n=1}^{N} P(n)\int P(\mathbf{x}|n)\log P(\mathbf{x}|n)dx \\ &\quad +\sum_{n=1}^{N} P(n)\int P(\mathbf{x}|n)\log P(\mathbf{x})dx. && (4.8) \end{aligned}$$

The expression for V is drastically simplified when put in terms of entropy

$$H(x) = -\int_{-\infty}^{+\infty} P(x)\log(P(x))dx. \tag{4.9}$$

The total overlap cost, in terms of entropy, becomes

$$V = -H(\mathbf{x}) + \sum_{n=1}^{N} P(n)H(\mathbf{x}|n) \tag{4.10}$$

$$= -H(\mathbf{x}) + \frac{1}{N}\sum_{n=1}^{N} H(\mathbf{x}|n). \tag{4.11}$$

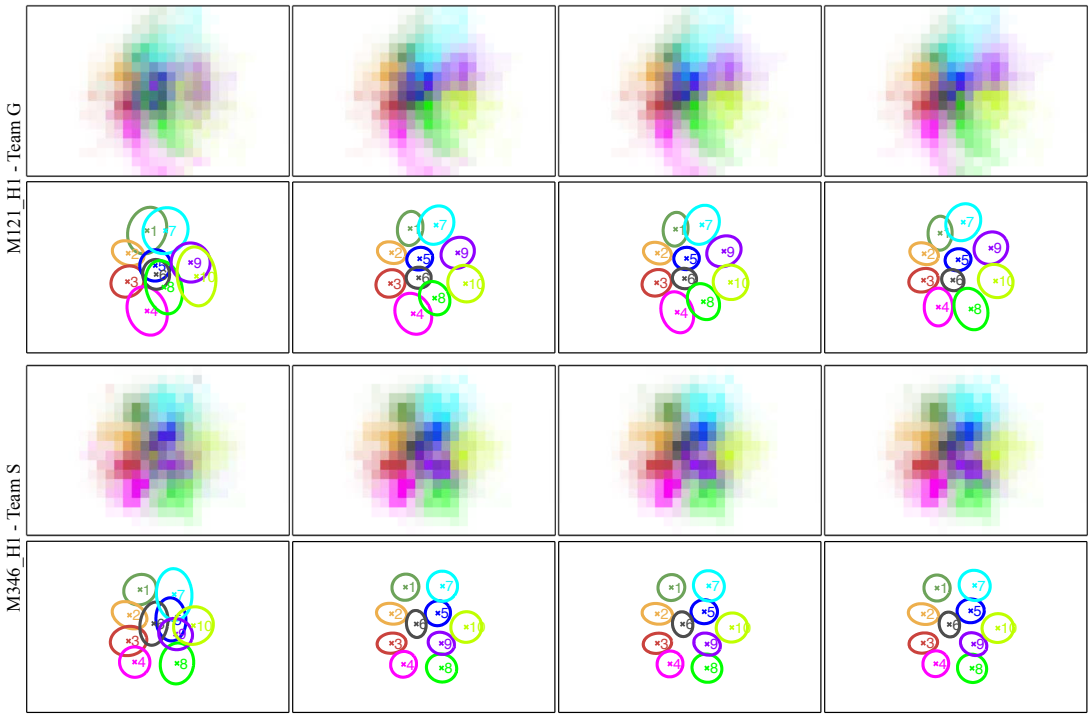
Substituting (4.11) into (4.5) and ignoring the constant term $H(\mathbf{x})$, the optimal formation is the set of role-specific probability density functions that minimise the total entropy

$$\mathcal{F}^* = \arg\min_{\mathcal{F}} \sum_{n=1}^{N} H(\mathbf{x}|n). \tag{4.12}$$

4.3.1 Procedure

As there is no way to solve this problem efficiently, an approximate solution can be achieved using the expectation maximisation (EM) algorithm [38]. The proposed procedure is similar to k-means clustering except with the constraint that at each frame, each agent (or player) must be assigned to a unique role. Instead of assigning each data point to its closest cluster, the linear assignment cost of assigning roles to agents (players) is minimised at each frame using the Hungarian algorithm [79], to ensure there is a one-to-one assignment of roles to agents (players).

The procedure is described as follows and is visually presented on sports data in Figure 4.2 for two match halves. Firstly, the data is normalised so that teams are attacking from left to right and the effects of translation are negated by setting the



(a) Initial Roles (b) Iteration 1 (c) Iteration 2 ... (d) Final Roles

Figure 4.2: Example of the role discovery procedure for two teams, showing the role distributions at each iteration (drawn as heat maps in the top row and 2D Gaussians in the bottom). The initial role distributions (a), are calculated by assuming each player is assigned a single role over all frames and taking their distribution over the half. A high degree of overlap is exhibited due to frequent positional swaps between players. Taking (a) as the template, each frame is assigned to these roles and the updated distributions are shown in (b). This is then used as the template for the next iteration and the procedure is repeated until convergence, resulting in well separated role distributions as in (d).

tracking data to have zero mean in each frame. This results in a formation being represented as the spatial distribution of each role relative to the team's centroid. The initial formation is set by arbitrarily assigning each player a unique role label at the start of the match and maintaining these roles throughout the entire duration of the tracking data. Even though there is heavy overlap between the distributions of some players, initialising based on player identity is a reasonable estimate of the formation as it is assumed that players tend to play one role for the majority of the time. Examples of the initial occupancy maps for each role are shown in Figure 4.2 (a). Role labels are then assigned to player positions at

each frame of the tracking data by formulating a cost matrix based on the log probability of each position being assigned a particular role label. The Hungarian algorithm [79] is used to compute the optimal assignment of role labels based on the current formation template. Once role labels have been assigned to all frames of the tracking data, the probability density functions of each role are recomputed, giving an updated formation template. The process is repeated until convergence, resulting in well separated probability density functions as in Figure 4.2 (d). In this way, each player is assigned to a role at each frame of the tracking data and the role probability distributions ($P_n(\mathbf{x})$) are discovered, providing the formation that the team played over the match half.

4.4 Individual and Team Analysis

The proposed formation discovery procedure was performed for each team and match half, excluding formations where players were sent off, resulting in the detection of 1411 formations. Each formation consists of a set of ten distinct role probability distributions, representing the structural arrangement of the team over a half, and depicts the long-term characteristic behaviour of the team.

4.4.1 Visualising Team Formations

The formations for each of the 20 teams (A-T) for every match are shown in Figure 4.3. As can be seen in this figure, most of the teams tend to play the same formation across the season with only a slight variation occurring in some of the positions. For example, only teams B and T seem to have some variation across the course of a season, while others like teams A, F, P and R only have a minor change in the midfield (i.e. one holding midfielder vs two, or playing with one

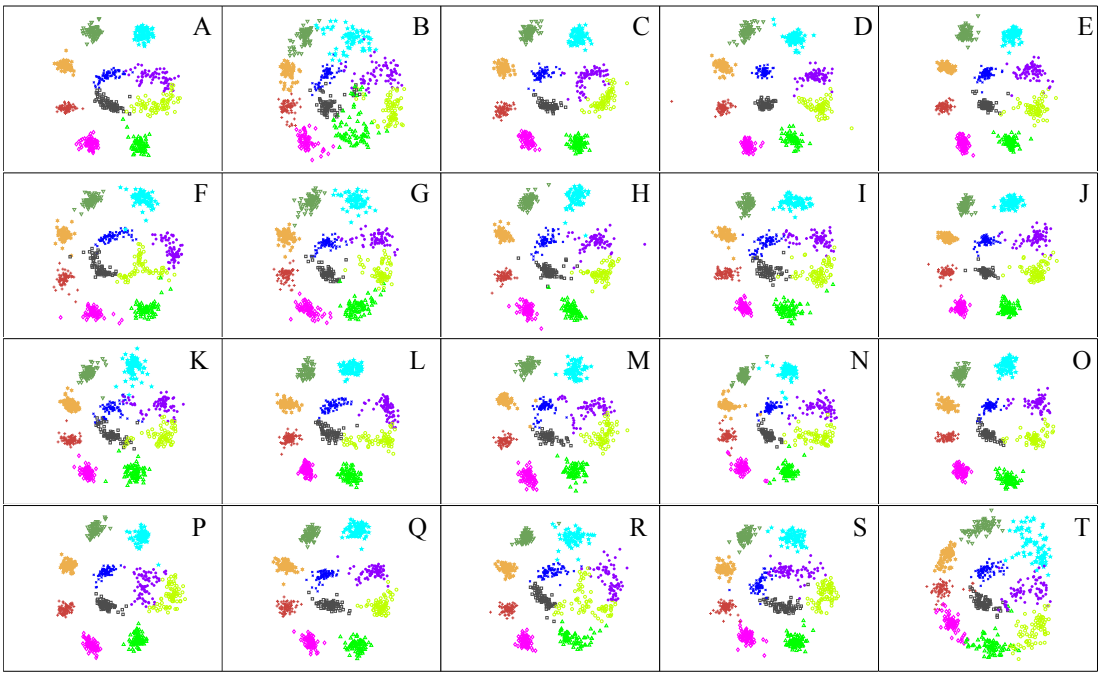


Figure 4.3: The discovered formation descriptors are displayed for each team (A to T). The formations are drawn so that teams are attacking from left to right, with colours representing different roles. For clarity of visualisation, only the centroid for each role for each match is shown instead of displaying the full distribution.

striker vs two). Other than that, most teams tend to be rather staunch in what they play. The most dominant formation appears to be a 4-4-2, with some teams varying the midfield as described above. Only one team appeared to play with three defenders (team T).

In addition to representing the long-term behaviour of the team in terms of formation or team structure, the proposed method can also be used over shorter durations to dynamically represent how a team plays throughout a match. Compared to existing statistics which only contain sparse team information (e.g. # corners, # shots, % possession), the proposed approach can represent the spatio-temporal characteristics of the match in terms of formations and position. One of the statistics which broadcasters present during a live-broadcast is the possession duration of both teams over the past 5 minutes which gives an indication of which

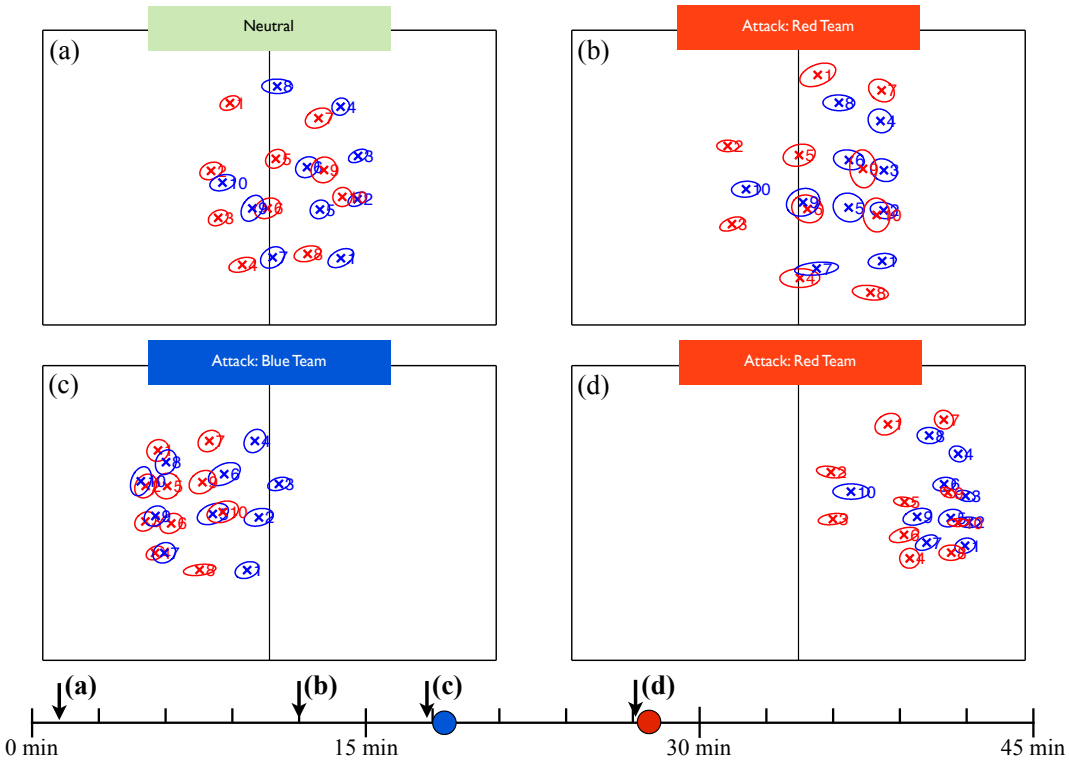


Figure 4.4: Film strip representing the timeline of a match in terms of formation. By observing formations within a match using a sliding window of 5 mins, the game progresses in terms of team structure and location on the field can be seen. A 45 min half timeline is shown (circles = goals) with the home team (red) attacking from left to right. (a) During a neutral portion of the game, it can be seen that both teams are playing a 4-2-3-1 formation. (b) Next, it can be seen that the red team makes an attack by spreading out and advancing its players forward. (c) Before the blue team scores, the centre midfielder (role 9) moves forward to aid in the attack. (d) In the final example, the red team scores, with the whole team positioned close to the goal.

team is dominating. While this is insightful, it does not give any information about where this is happening. Using a sliding window of 5 minutes on the role assigned player positions, the play progression can be visualised in terms of team formations and relative player positions, by using 2D Gaussians to represent the role distributions over the time window. A film-strip of this approach is shown in Fig. 4.4.

4.4.2 Clustering Team Formations

To get an indication of the types of formations used by teams across the league of data, agglomerative clustering was employed on the formations. In agglomerative clustering, each observation starts in its own cluster and pairs of clusters are merged based on distance, forming a cluster hierarchy. The distance between formations was calculated as the sum of the Earth Mover's Distance (EMD) [121] between corresponding role probability density functions. Agglomerative clustering was chosen as it provides a flexible and non-parametric approach to discover the types of formations used across the dataset. Different clustering thresholds of the hierarchy can be observed, and a cut-off of six clusters is shown in Fig. 4.5. Six clusters were chosen as this allows the coarse categories of formations to be visualised. Segregating further resulted in clusters that look very similar, while a smaller number had too much variation within the clusters. From Fig. 4.5, it can be seen that clustering resulted in the discovery of distinct formation classes - e.g. Cluster 2 and 3 have only 1 striker in the front, Cluster 1 and 5 have 2 strikers, while Cluster 4 and 6 appear to have 3. Cluster 4 is the only cluster with 3 defenders at the back with the remainder all having 4.

By observing the clustering assignment frequency (top right of each cluster in Fig. 4.5), an indication of which formations are more commonly adopted by teams can be seen. Cluster 1, which appears to be a 4-4-2, is the most common formation with approximately 54.11% of formations being assigned to this cluster, followed by Cluster 2 (22.30%), which appears to be a 4-2-3-1. This gives insight into the strategies adopted by teams (e.g. having 2 strikers instead of 1 may be considered a more attacking strategy).

The clustering results were evaluated by comparing the cluster groups against ground truth formation labels. The ground truth labels were annotated by a

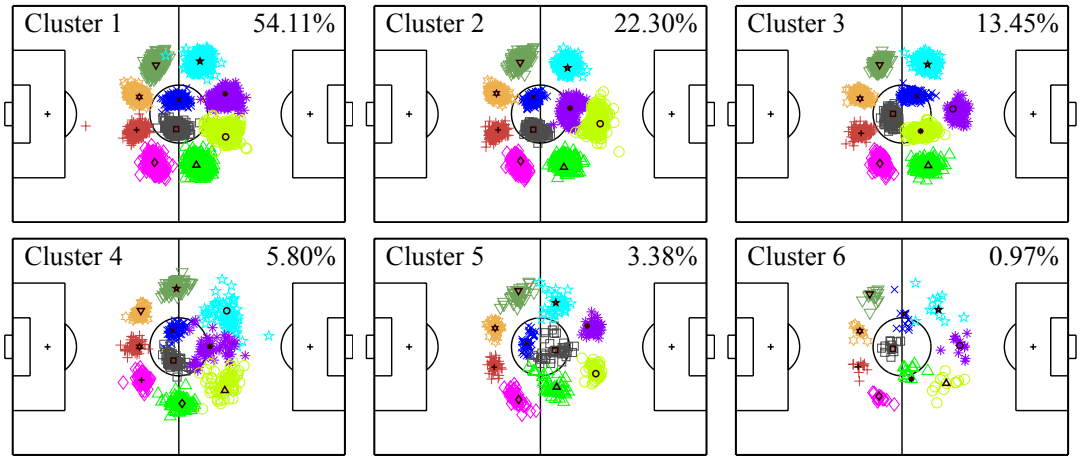


Figure 4.5: Formation clustering output. The formations assigned to each cluster are shown with the median formation overlaid in black. Each dot point represents the discovered mean role position for the formation for a match half, and the percentages refer to the proportion of examples assigned to each cluster, indicating the popular formations adopted by teams across the league. All formations are normalised so that the team is attacking from left to right.

soccer expert who annotated the most frequently observed formation for each match half and each team according to the arrangement of players in defensive, midfield and attacking lines (4-4-2, 4-2-3-1, 4-3-3, 3-4-3, 4-1-4-1, or 'other' where the team either did not display a dominant formation or was not one of the given labels). To evaluate the results, the label of each cluster was estimated as the most frequent ground truth label within the cluster and the results are presented as a confusion matrix in Fig. 4.6.

It can be seen from Fig. 4.6 that the discovered formation clusters match the ground truth annotations quite well, with high within cluster label agreement and an overall correct classification rate of 75.33%. The most confusion is in Cluster 5 which appears to be a 4-1-3-2 formation (referred to as a 4-4-2 'diamond'), often being classified as a 4-4-2 and 4-3-3. On visual inspection of the misclassified examples, sometimes the formation appears in between two clusters, and there is some confusion between the 4-4-2 and 4-2-3-1 formations when the second striker

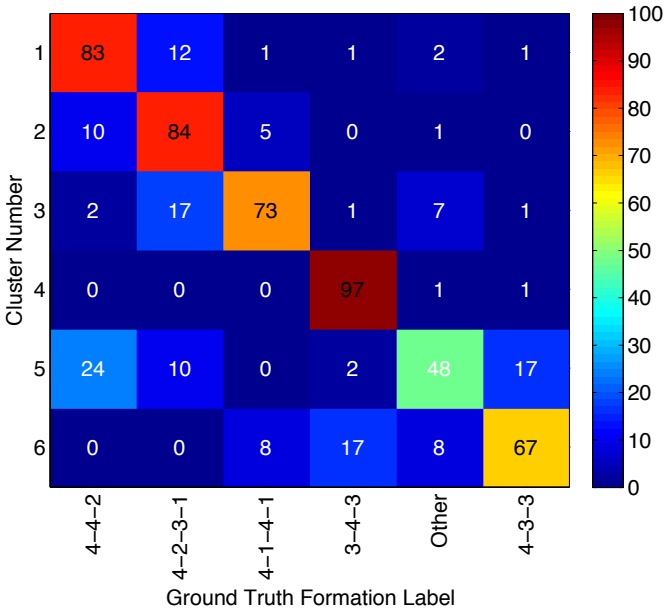


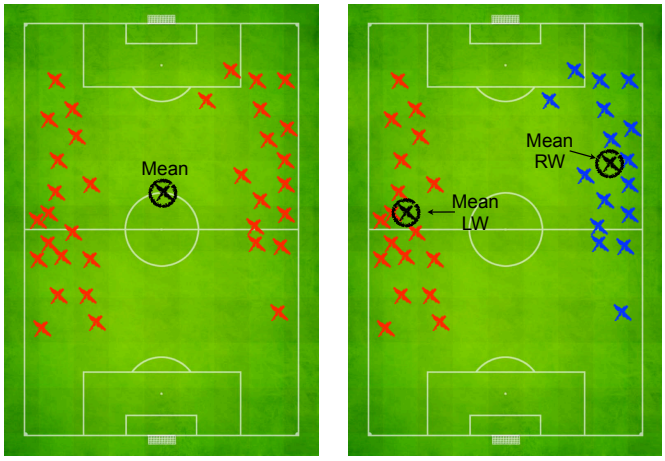
Figure 4.6: Formation clustering results presented as a confusion matrix, showing the proportion of each cluster belonging to each ground truth formation label.

is positioned slightly behind the other.

4.4.3 Individual Player Analysis

Compared to existing analysis which often only looks at the mean behaviours of each player, the role assignment method dynamically assigns players to roles throughout a match and therefore allows the different characteristic behaviours of each player to be analysed and visualised. An example of where this is useful is shown in Figure 4.7.

The roles of each player over a match half relative to the discovered formation are first examined, as shown in Fig. 4.8. In these examples, the behaviour of two teams is shown, demonstrating how players dynamically alternate positions throughout a match and how versatile they are within the formation. Plot (c) represents a 5 min smoothed version of the role assignments (to ignore temporary



(a) (b)

Figure 4.7: Roles provide important context for performing individual player analysis. (a) Shows the touches of the player who starts in the left-wing position but changes half way through the half to the right-wing. Current approaches just give the mean position which neglects the context. (b) Using the proposed formation and role-representation captures context to allow for better individual player analysis.

role swaps) showing dominant roles taken by each player. From this, it can be seen that in the top game, roles remain constant throughout the match, while in the second game the midfielders (roles 5 and 6, shown in blue and grey) alternate positions frequently throughout the match. The most frequent role swap are visualised as a transition matrix in plot (d), which also gives an indication of team playing style.

Next, it is demonstrated how roles can be used to provide context in analysing player events throughout a match. In Figure 4.9, all the events that occurred within an example match half are displayed. On the left the events were segmented by role, and on the right the events were segmented by player identity. In (b), interesting behaviour can be observed for the players playing left wing and right wing who swap roles for part of the match. The role representation is able to detect these characteristic behaviours (coloured in green and cyan). If the mean of each player's actions were simply taken, this important tactical variation would be missed. Roles provide important context for such player analysis.

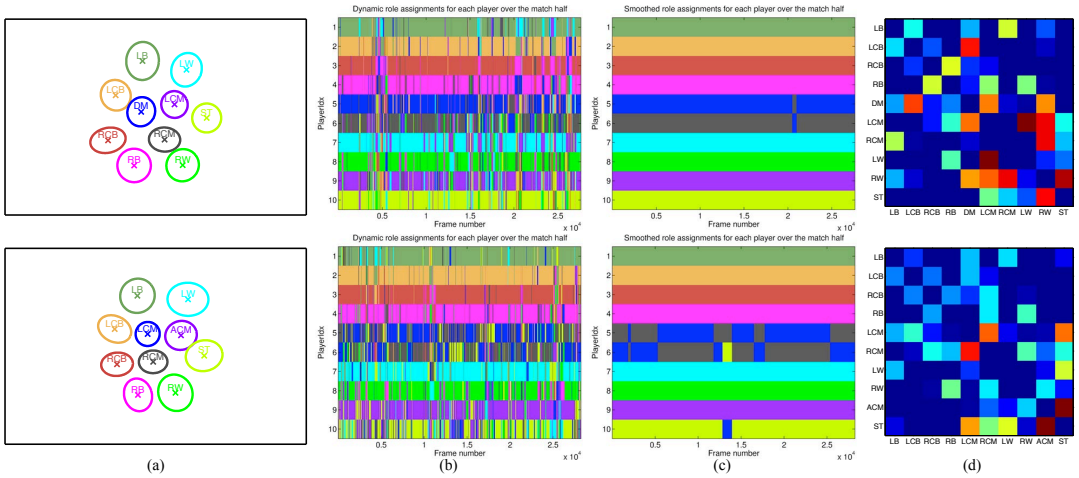


Figure 4.8: The behaviour of two different teams over half a match, demonstrating: (a) Their overall formation calculated using the proposed formation discovery procedure (with roles represented as 2D Gaussians). (b) A timeline showing the role assigned to each player at each frame, coloured by role. (c) A 5 min smoothed version of the role assignments (ignores temporary role swaps), (d) The per-frame role swaps across the half {left-back(LB), left-center-back(LCB), right-center-back(RCB), right-back(RB), left-centre-midfield(LCM), defensive-midfield(DM), right-centre-midfield(RCM), left wing(LW), right-wing(RW), attacking-center-midfielder(ACM), striker(ST)}

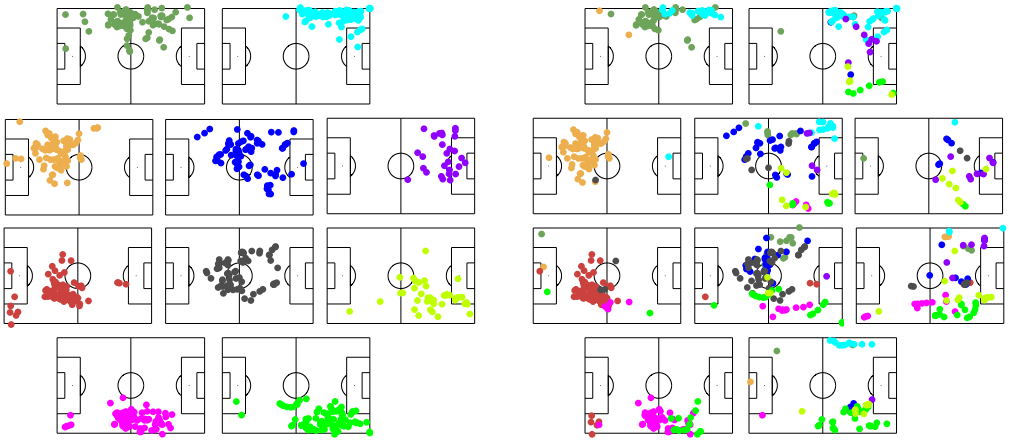


Figure 4.9: Every event within a match half segmented into (a) roles, versus (b) player identity (both coloured by the role of the player at the frame of the event)

4.5 Predicting Team Identity

To determine if teams had a distinct playing style, a series of team identity experiments were conducted. The challenge was, *given only player tracking data and ball events, how can the identity of each team best be predicted*? To do this,

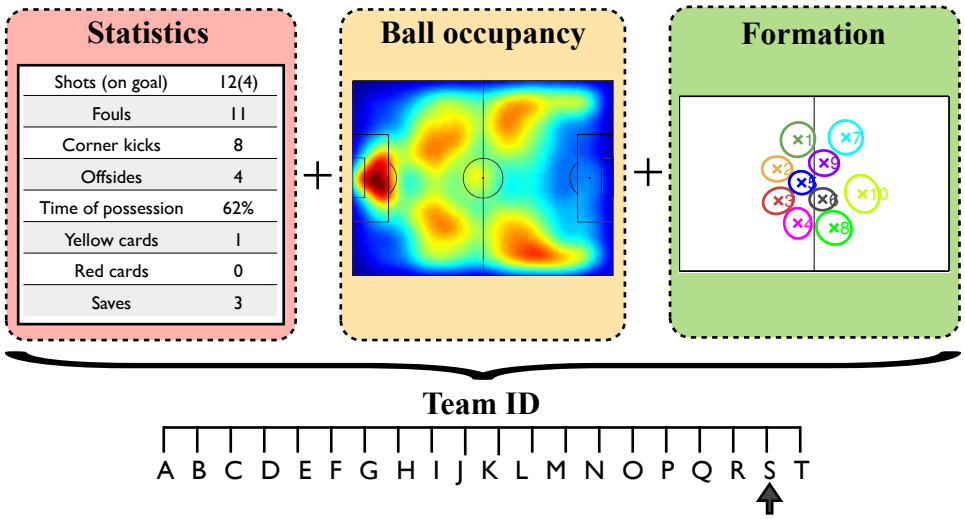


Figure 4.10: Based solely on match statistics, ball movement patterns, and the formation descriptor, the identity of a soccer team can be predicted.

three types of match descriptors which describe team behaviour were generated: 1) match statistics, 2) ball occupancy, and 3) team formation, and these are used for predicting team identity as shown in Figure 4.10.

4.5.1 Match Descriptors

Match Statistics: During a match, various statistics that capture team and individual behaviour are annotated. Table 4.2 presented at the start of the chapter, lists the statistics that were annotated and used for representing team performance. While the number of these match statistics is quite large, the majority are quite sparse with only a couple of these events labelled per match. Only a half-dozen of the most important match statistics are normally documented in reporting of a match (i.e. goals, shots on target, shots off target, passes, corners, yellow and red-cards).

Ball Occupancy: Associated with the match statistics/events are the time and location for each occurrence. To form a representation of this information, the

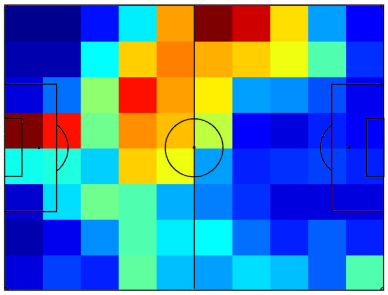


Figure 4.11: Example of a quantised ball occupancy map (10×8 grid) of a team from a match (attacking left to right).

approach used in [94, 96] was adopted which consists of estimating the continuous ball trajectory at each time-stamp by linearly interpolated between events, as well as which team had possession (ignoring stoppages). The field was split into a 10×8 spatial grid and ball occupancy of each of these grids for each team were calculated (i.e. how often the team was in possession of the ball in this location over the match). A visualisation of a ball occupancy example is shown in Figure 4.11.

Formation Descriptor: For each match half, the formation descriptor $\mathcal{F}^*$ was found using the method described in Section 4.3.1. This gave an $M \times N$ matrix where M refers to the number of cells in the field and N is the number of roles (set to 10, as the goal-keeper was omitted, as well as games which had a player sent off). A depiction of the formation descriptors for each team for all matches was presented in Figure 4.3. As teams are rather rigid in the way they play across a season, it suggests that this is a useful feature in discriminating between different teams. Another interesting point is, as teams vary little in terms of playing style throughout the season, this could be used as a powerful prior for preparing against an opposition in upcoming matches.

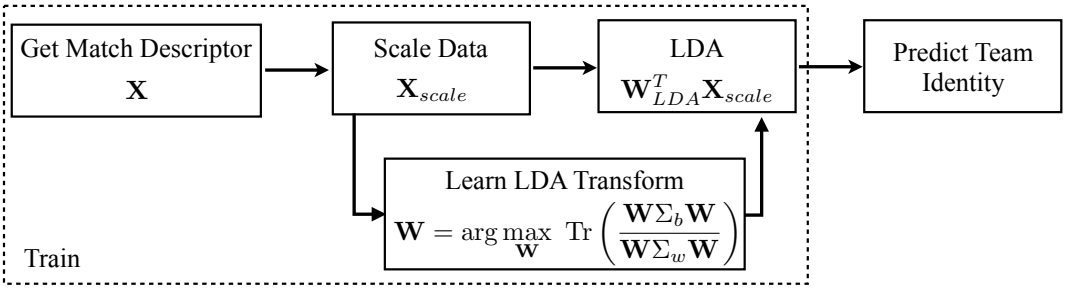


Figure 4.12: Block diagram for learning the discriminative feature vector and predicting team identity. Given a match descriptor, the data is first scaled then multiplied by $\mathbf{W}^T$, found using LDA, to yield a discriminative feature vector. The LDA matrix is learnt using the team identity labels and their match descriptors in the training set. Team identity is then predicted using k-NN.

4.5.2 Experiments

Experiments were performed to determine which features were best at discriminating between teams and which best characterise a team's behaviour. The team identity experiments were performed using a "leave-one-match-out" cross-validation strategy where one match was left out to test against, and the remaining matches were used as the train set. A block-diagram shown in Figure 4.12 describes the procedure.

To begin, the three descriptors described in the previous section were generated and were linearly scaled to be in the range $[0, 1]$, to ensure equal weighting of features. To obtain a compact but discriminative representation, linear discriminant analysis (LDA) was used to learn the transformation matrix $\mathbf{W}$ from the training set, using the team identity as the class labels (i.e. $C = 20$). LDA was chosen as it explicitly models the difference between classes and helps to determine the distinguishing features of a team. After learning the $\mathbf{W}$ matrix, the features were then multiplied by $\mathbf{W}^T$ to yield a lower dimensionality discriminant feature vector of dimensionality $C - 1$. To predict the identity label of the teams in the test match, a k-nearest-neighbour classifier was used with the Euclidean norm as the

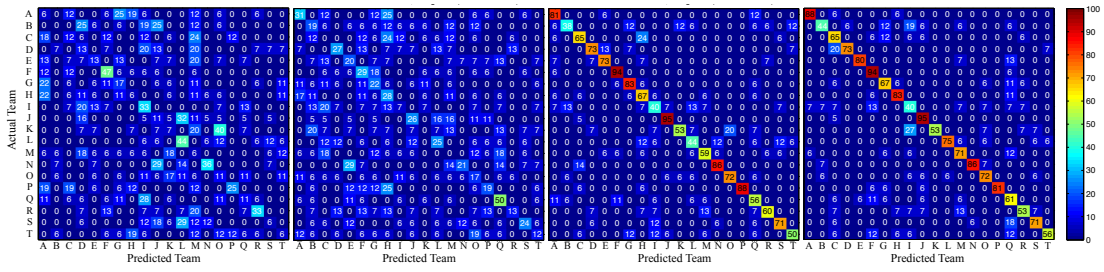
distance metric. A neighbourhood of $k = 20$ was chosen as this provided the best results for most descriptors, however, the order in performance of the different descriptors was consistent across various k.

The results for the various descriptors are shown in Figure 4.13. In Figure 4.13(a), it can be seen that using only match statistics is a poor indication of team identity with an overall accuracy of 17% (chance is 5%). This result makes sense as the match statistics only contain coarse event information without any spatial or temporal information about the ball or the players. Using the ball occupancy gave marginally improved performance over the match statistics with an accuracy of 19% (Figure 4.13(b)). This is well below the 33% which was obtained in the previous works [94, 96]. A possible explanation of the performance difference could be due to the coarse estimation of the possession strings and the ball occupancy maps from the event data.

The most impressive performance by far is the formation descriptor which obtains over 67% accuracy (Figure 4.13(c)), which clearly shows that teams have a true underlying signal which can be encapsulated in the way the team moves in formation over time. The descriptors were also fused together by concatenating all the scaled features and this approach improved the overall performance to over 70% (Figure 4.13(d)) which shows that there is complimentary information within the other descriptors. A bar-graph comparing the overall performance for each descriptor is given in Figure 4.14.

4.6 Analysing Team Style

Team style is a very subjective and difficult attribute to label, especially in continuous sports like soccer. This is in part due to the dynamic and low-scoring



(a) (b) (c) (d)

Figure 4.13: Team identity results for the various descriptors: (a) match statistics, (b) ball occupancy, (c) formation descriptor and (d) fused all descriptors.

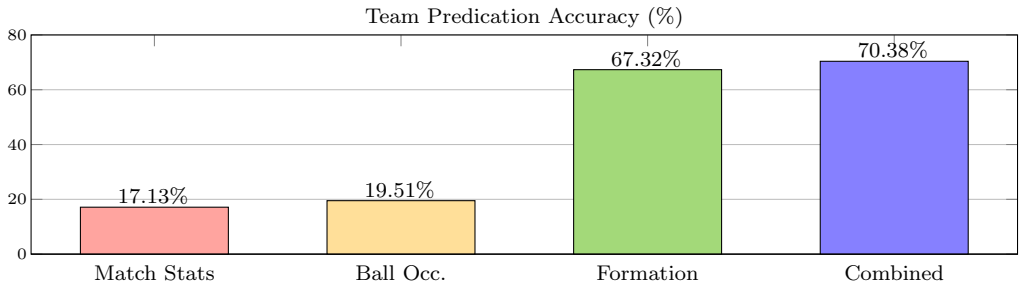


Figure 4.14: Comparison of the team identity prediction accuracy for different descriptors.

nature of such sports which makes it difficult to segment the game into discrete parts, as well as the fact that style is a high-level attribute that encompasses all aspects of play, from how a team scores, to how they cover the field, and how they interact and move in co-ordination. The formation descriptor incorporates many of these aspects and is used in this section to approximate team style for prediction and anomaly detection tasks.

4.6.1 Team Style

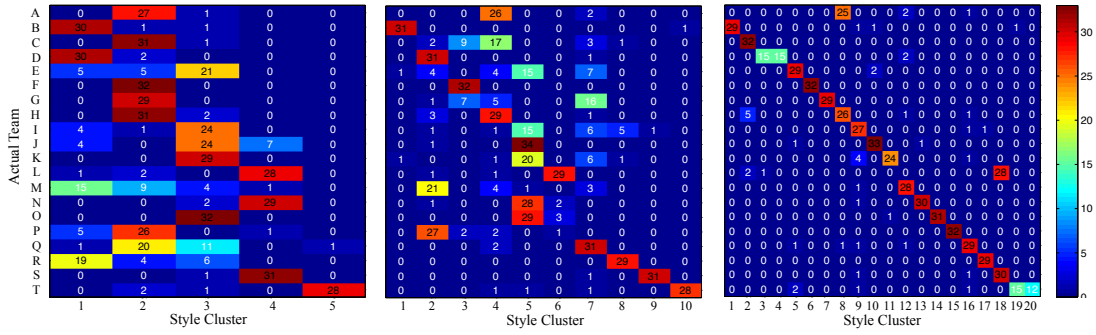
To evaluate team style prediction and anomaly detection, the soccer data was split into separate training and testing sets. The last two rounds of the season were excluded for testing, and the remaining games were used to train the style models. This provides a realistic chronological evaluation framework where analysts may

want to predict future performances, while providing sufficient data to train off.

Teams only verse each other twice throughout the season, resulting in insufficient data to model team versus team behaviours. Instead, a discrete set of styles is learnt through clustering to provide more examples of behaviours for each style. Given a training set of team behaviour descriptors, a discrete set of styles was learnt using k-means clustering. The clustering was performed on the match features projected into the lower dimensionality discriminative space using LDA as in the team identity experiments (Fig. 4.12), as this provides better discrimination between styles and faster prediction. K-Means clustering was adopted as it separates the data into classes of relatively equal proportions, providing sufficient examples of each style.

Style clustering results for $k = 5$, 10 and 20, are shown in Figure 4.15. It can be observed that there is some overlap in style between certain teams, and some teams exhibit multiple styles. The variation in style for each team across the season using $k = 5$ styles, is shown in Figure 4.16. Team T stands out, being in a style cluster of its own, which could be explained by the distinctly different formation from all other teams, with 3 defenders at the back (as was discovered previously in Fig. 4.3). Most teams play a single style, while teams E and R vary their playing styles more frequently than other teams.

To encapsulate the behaviour styles that teams adopt, the playing style of a team is defined as a linear combination of their styles in previous matches. A style vector is constructed using the normalised weights from the style clustering matrices. For example, using the 5 style clusters from Figure 4.15(a), the style vector for Team A=$[0, \frac{27}{28}, \frac{1}{28}, 0, 0]$ and Team B=$[\frac{30}{32}, \frac{1}{32}, \frac{1}{32}, 0, 0]$. Modelling teams as a combination of the styles they play makes intuitive sense, as sometimes a team could play a pressing game and on other occasions the team may play defensively, so they would be weighted according to these performances. Another



(a) (b) (c)

Figure 4.15: Results for clustering the descriptors of each match half when setting the number of style clusters to: (a) 5, (b) 10, and (c) 20.

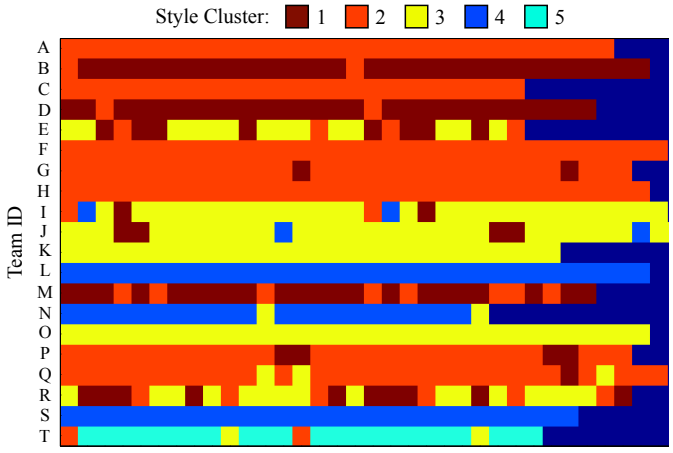


Figure 4.16: Shows the variation in style each team has across a season when 5 style clusters are used. Each coloured block represents the formation style the team played for a match half and they are concatenated chronologically, excluding match halves that were missing data or had a player sent off (i.e. < 10 field players).

team may be very rigid and play the same style every game – so the weight for that style would be very high. These style vectors can then be used to assist in prediction of future behaviours.

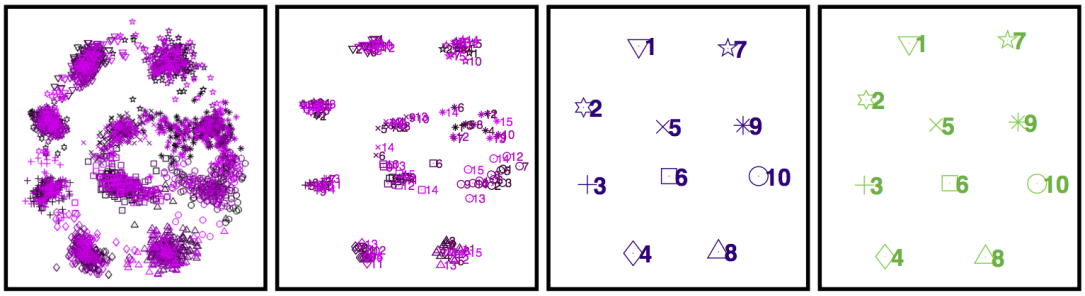


Figure 4.17: Formation prediction procedure using k-NN regression. (a) all training examples, (b) retrieved examples according to style prior, (c) the predicted formation (= mean(retrieved examples)), (d) the actual formation.

4.6.2 Prediction and Anomaly Detection

Previously, given the ball and player tracking data, team identity was predicted. In this section, the reverse is done - *given only the identity of the two teams playing, the playing styles of the two teams are predicted.*

To predict the most likely features, K-nearest neighbour regression was used. While a more complicated regressor could be used, this simple model demonstrates the utility of the formation descriptor in describing and predicting team style. Given two team names, their team style vectors learnt from the training data were used to predict the most likely formation they will adopt in the coming match, by regressing from the most similar training examples. That is, for each match in the training set, the two team styles are compared to the test match's team style priors, and the K most similar matches in terms of team styles are then extracted. The mean of these features is then used to predict the features in the test match. The formation prediction procedure is shown in Figure 4.17.

The last two rounds of the season (containing 18 matches) were used to evaluate the prediction of team formation. The performance was evaluated by comparing the predicted formation to the actual formation played for each match, and the results are presented as the average error for each role position, presented in

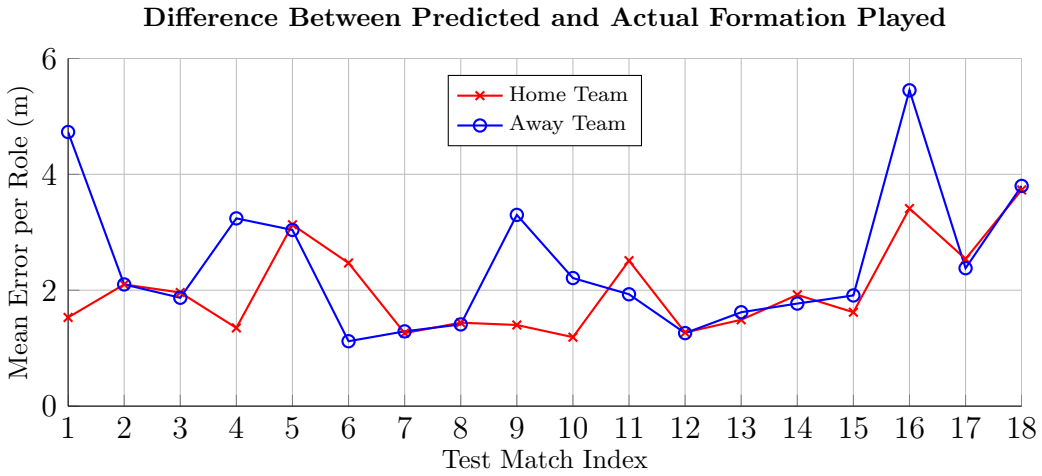


Figure 4.18: Results comparing the predicted formation to the actual formation played for the home (red) and away team (blue) for each match in the final two rounds of the season (18 matches) . The values refer to the average error for each role position.

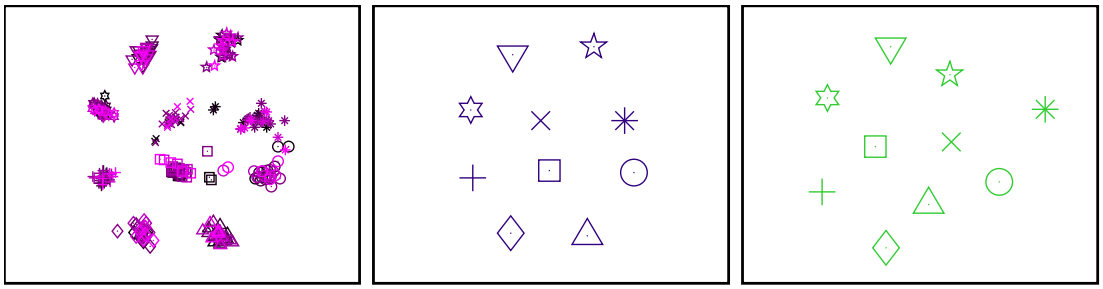


Figure 4.19: Example of a poor formation estimate (test match 16), which appears to be due to an anomaly in the team's behaviour. (a) retrieved examples, (b) predicted formation, (c) actual formation

Figure 4.18. It can be seen that most matches are estimated within 2 m average error per role, while Match 1 and 16 are most poorly estimated. This suggests that the teams were not playing their normal formation style in these matches and suggests anomalous behaviour. The predictions allow the most likely formation to be visualised and can also be used to detect anomalies, in which the predicted behaviour is very different from the observed formation, such as in Figure 4.19.

4.7 Exploring the Home Advantage

In terms of analysing soccer matches, two of the most important factors to consider are: 1) the formation the team played (e.g. 4-4-2, 4-2-3-1, 3-5-2 etc.), and 2) the manner in which they executed it (e.g. conservative - sitting deep, or aggressive - pressing high). Despite the existence of ball and player tracking data, such analysis is still performed manually by humans and the method proposed in this chapter is the first to automatically detect and visualise formations. In this section, the utility of the approach is showcased by investigating the "home advantage" and exploring whether there is any strategic reasons for why home teams are much more likely to win compared to away teams.

4.7.1 Statistics Highlighting the Home Advantage

In recent work, it was shown that home teams had significantly more possession in the forward-third which correlated with more shots and goals while the shooting and passing proficiencies were the same [96]. While teams had approximately the same number of passes, passing accuracy and shooting accuracy when playing at home and away, there were significantly more shots and goals for home teams, and it was found that this coincided with home teams having more possession in the forward third. Before proceeding, the same experiments were performed on this soccer dataset (details were given in Section 4.2), to see if the same phenomenon occurred. The comparison of home and away statistics are displayed in Table 4.3.

From Table 4.3, it can be seen that the same pattern exists in the data set used for this work. Most notably, home teams had more shots and goals, as well as more points and possession in the forward third, while the shooting proficiency and number of passes were roughly the same. This suggests something is different

Event Statistic	Mean for Home Team	Mean for Away Team	P-Value
Points	1.6	1.1	< 0.001
Goals	1.6	1.2	< 0.001
Shots on target	6.5	5.2	< 0.001
Shots not on target	8.9	7.0	< 0.001
Shooting accuracy	41.7%	42.4%	0.5046
Number of passes	451	436	0.1483
Possession in final third of field (of time in play)	14.1%	11.8%	< 0.001

Table 4.3: Mean match statistics highlighting the home advantage. A low p-value indicates that the null hypothesis should be rejected (that there is no difference in the home and away team distributions for the given statistic).

in the way the teams play: does it mean that teams play with two strikers at home while they play with an extra midfielder away, or is this a global trend of the formation where all the players pushed forward? To do this analysis, the team formations were detected and visualised.

The formations detected through the procedure described in Section 4.3.1, are now coloured for each match by whether the formation was played at home or away, as presented in Figure 4.20, with red corresponding to home performances and blue corresponding to away. It can be seen that most of the teams tend to play the same formation regardless of whether they are playing at home or away with only a slight variation occurring in some of the positions.

Given that teams tend to play a similar formation regardless of whether they play at home or away, this led the analysis to exploring *how* the formation was played (i.e. were they more attacking, or were they more defensive?). To answer these questions, the centroid of each formation was examined for each team when they were: a) in possession, and b) when the opponent was in possession of the ball. The times when the ball was out for a stoppage were disregarded, which is amazingly around 50% of the time. These plots are presented for the zoomed

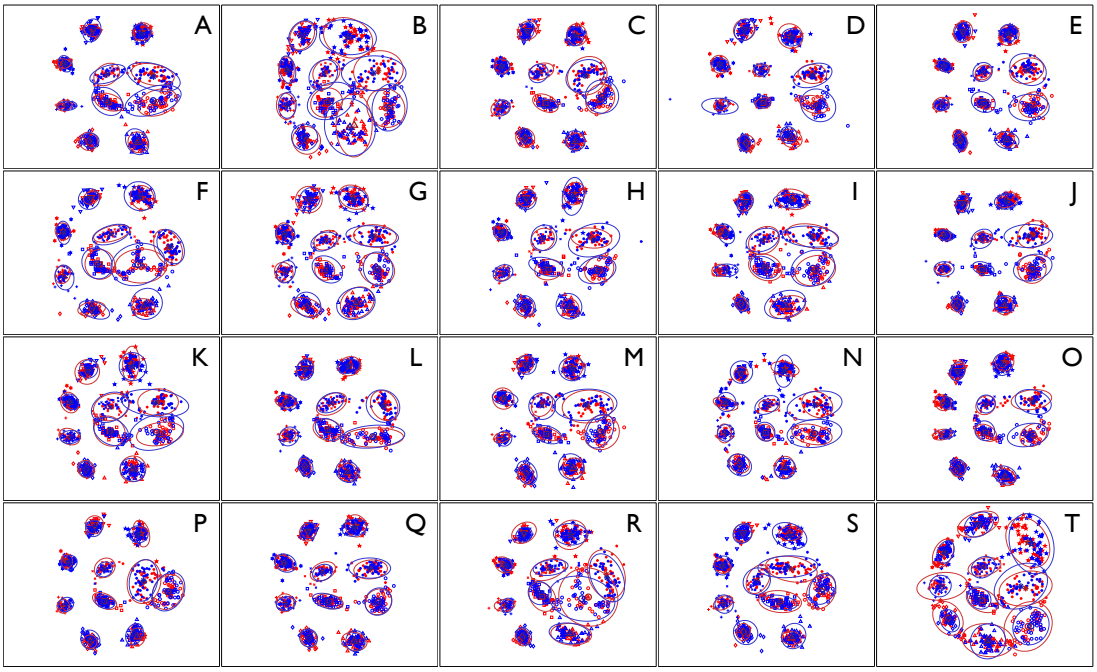


Figure 4.20: Formations for each team (A to T) comparing home (red) and away formations (blue), drawn so that teams are attacking from left to right. The mean position for each role is plotted for each match and the distribution of the mean role positions are drawn with ellipses for home and away games.

in area of the field (see Figure 4.21) in Figure 4.22 and 4.22 respectively. It is apparent from these plots, that nearly all teams are positioned further forward at home than they are away, both when attacking and when they defend. This can help explain how teams have more possession in the forward third at home, as simply having the players in more advanced positions suggest that they would have more possession in these advanced areas. Similarly, when they are defending higher up the pitch, it means that they are more likely to regain possession higher up the pitch. The downside to this tactic is that they leave more space exposed behind the defensive line which allows teams to be hit on the counter attack. However, in addition to winning the ball in more advanced positions (which is closer to the opponent's goal and is likely to lead to more shots and more goals), the home team will expend less energy. This has large implications as the less energy a team expends, the longer and more sustained attacks a team can lead.

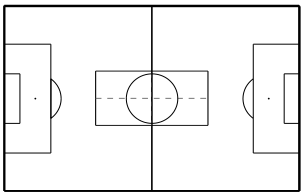


Figure 4.21: To get a closer look at the formation differences, analysis was conducted on a zoomed in area of the field.

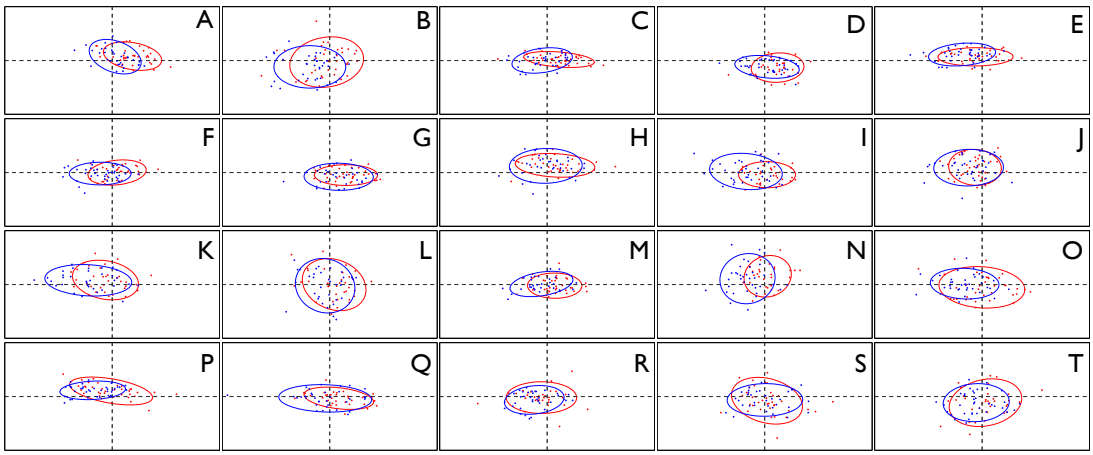


Figure 4.22: Mean position of the team when they were in possession. Each dot represents the centroid for a match half and the ellipses denote the distribution of the team's home games (in red), and away games (in blue).

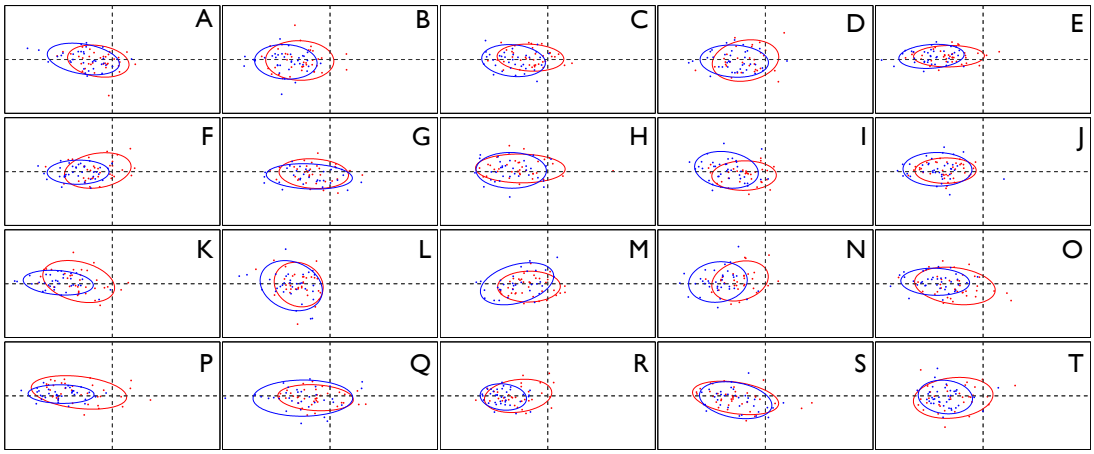


Figure 4.23: Mean position of the team when the opposition was in possession. Each dot represents the centroid for a match half and the ellipses denote the distribution of the team's home games (in red), and away games (in blue)

4.8 Summary

In this chapter, a method was proposed to align multi-agent data in an unsupervised manner, by minimizing the entropy of a set of role distributions. An Expectation Maximisation approach was proposed to efficiently solve this problem, which simultaneously assigns agents of a group (players) to spatial roles at each frame of the dataset and discovers the overall group (team) structure or formation. In this procedure, the variance in location of each role is reduced, disentangling the agent or player distributions into distinct role distributions to allow the discovery of the underlying group structure. Compared to the method proposed in Chapter 3 which required ground truth labels to learn the formation representation, the proposed method is completely automatic and learns the formation of a group directly from data.

The alignment of the data enables a host of new group behaviour analysis tasks to be performed such as discovery and visualisation of group structure, as well as improved clustering, prediction and group identity classification. The utility of the approach was demonstrated in performing large-scale individual and team analysis using a full season of data from men's professional soccer, consisting of over 21.5 million frames of player tracking data, spanning 20 teams and 374 matches. While the proposed approach is most suited to analysing team sports data, it can be applied without modification to any multi-agent data for which the dominant spatial structure across a period of time is desired and where the individual spatial positions of the agents may vary across that time (e.g. bird flight formations).

Using the formation discovery procedure, the team formations for each match half across a whole season of professional soccer were visualised and clustered to give an indication of group structure and strategy. The discovered formations were

then used to enhance individual player and team analysis, by providing context to the player statistics. The approach was also used to visually summarise a game which gives an indication of dominance and tactics. Following this, the formation descriptor's utility in discovering a low-dimensional discriminative feature space was demonstrated, by projecting the set of occupancy maps of each role into a lower dimensional space using linear discriminant analysis (LDA). The approach was shown to characterise individual team behaviour significantly better (3 times more) than other match descriptors typically used to describe team behaviour. A series of prediction and analysis tasks were conducted to highlight the application of the approach.

A case study was then performed on the home advantage phenomenon that exists in soccer, using the discovered team formations. It was found that while teams tend to play the same formation at home as they do away, the manner in which they execute the formation is significantly different. Specifically, it was shown that the position of the formation of teams at home is significantly higher up the field compared to when they play away. This conservative approach at away games suggests that coaches aim to win their home games and draw their away games. While enabling new discoveries of team behaviour which can enhance analysis, it is also worth mentioning that the automatic formation detection method is the first to be developed.

.

Chapter 5

Representing Noisy Data

5.1 Introduction

In many group behaviour analysis tasks, vision-based approaches to detecting and tracking people are preferable over wearable sensors as they can be acquired unobtrusively and do not require each individual to be instrumented. Unfortunately though, tracking people continuously over long time durations is still an unsolved vision problem due to challenges caused by occlusions, variations in resolution and pose, as well as strong illumination changes, resulting in the acquisition of noisy detection data. For tasks like clustering and retrieval, having noisy data (i.e. missing and false detections) is problematic as it generates discontinuities in the input data stream, and renders a lot of automatically acquired data useless without the time-consuming manual labour required to clean it up.

In this chapter, group context is used to perform analysis directly from noisy data in the sports domain. As player motion and position (i.e. proximity to team-mates and opponents) are heavily linked to game-context and where the

action on the field is taking place, these contextual features can be used to fill in the gaps of missed tracks caused by missed or false detections. An important contextual feature in team sports is characterised by a *formation*: a coarse spatial structure, defining a set of *roles* or individual responsibilities distributed amongst the players, which is maintained throughout the course of the match. Additionally, player movements are governed by physical limits, such as acceleration, which makes trajectories smooth over time. These two observations suggest significant correlation (and therefore redundancy) in the spatio-temporal signal of player movement data.

A core contribution presented in this chapter is the extraction of a low-dimensionality approximation of multi-agent time series location data. It is demonstrated how a *role representation* provides a more compact representation compared to player identity, and allows subspace methods such as the bilinear spatio-temporal basis model [3] to be used. The compact representation is critical for understanding team behaviour and enables the recovery of a true underlying signal from a set of noisy detections. This effectively allows the detections to be de-noised and allows for efficient clustering and retrieval of game events. To evaluate the approach, a fully instrumented field-hockey pitch consisting of 8 fixed high-definition (HD) cameras was utilised as well as a state-of-the-art real-time player detector, to provide approximately 200,000 frames of data on which the methods were evaluated and compared against manually labelled data.

5.2 Detection Data

5.2.1 Field-Hockey Test-Bed

To enable this research, a multi-camera test-bed at a professional field-hockey pitch was utilised to acquire detection data of group behaviours. The test-bed consists of eight stationary high definition (HD) cameras, attached to light fixtures, which together provide complete coverage of the 91.4 m $\times$ 55.0 m playing surface. An example image from each camera is displayed in Figure 5.1.

Figure 5.1: View of the field-hockey pitch from the 8 fixed HD cameras.

Using the test-bed, several matches from an international field-hockey tournament were recorded and analysed. The tournament consisted of 24 matches, played across two 30 minute halves and in this chapter, seven complete matches were analysed, totalling over 8 hours of match data for each camera captured at 30 frames per second. To process this vast amount of video frames, fast and robust computer vision algorithms had to be applied to convert the visual data into a suitable format to perform analysis.

5.2.2 Player Detection and Team Affiliation

For each camera, player image patches were extracted using a real-time person detector [28], which detects players by interpreting background subtraction results in terms of 3D geometry, where players are coarsely modelled as cylinders of height 1.8 m. This equates to a height of 40-100 pixels in the camera frames depending on the player's distance from the camera. To normalise for the scale variation, the image patches were normalised to a fixed size of 90×45 pixels, and were then classified into teams based on the colour histograms of their foreground pixels. The LAB colour space was used for representing the colours of each image patch, ignoring the luminance channel as it is affected by illumination changes. Nine bins were used for each dimension and the histograms were normalised to sum to one. The team classification procedure is illustrated in Figure 5.2

Models for the two teams were learnt using k-means clustering on a training set of approximately 4000 player histograms, and the Bhattacharyya coefficient was used as the distance metric. Each detected image patch was then classified to the closer of the two team models, or if it fell outside a threshold, it was classified into "others" which includes referees, goalies and false-positive player image patches. Since opposing teams wear contrasting colours in the tournament,

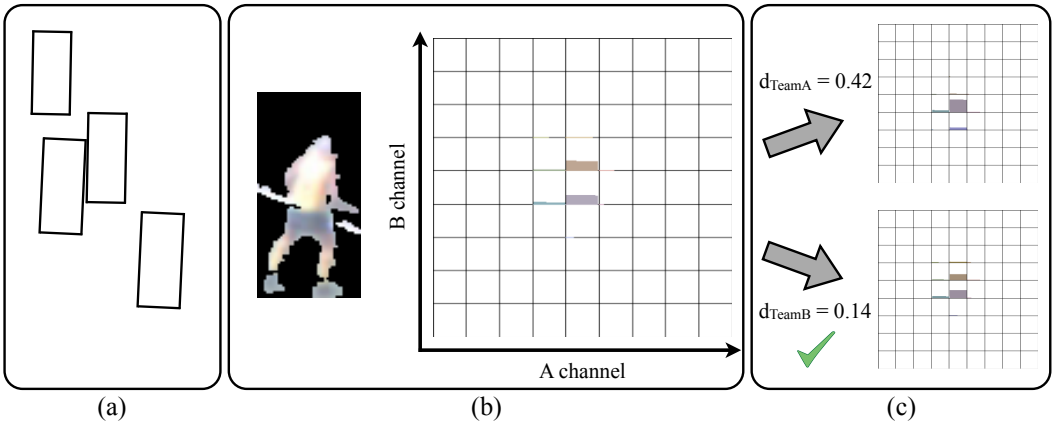


Figure 5.2: Team classification procedure. (a) Players were detected at each from using a real-time person detector. (b) The extracted image patches for each detection were then represented using a colour histogram of the foreground pixels in LAB colour space. (c) The image patches were then assigned to the closer of the two pre-trained team histogram models (or "other" if the distance to each model exceeded a threshold).

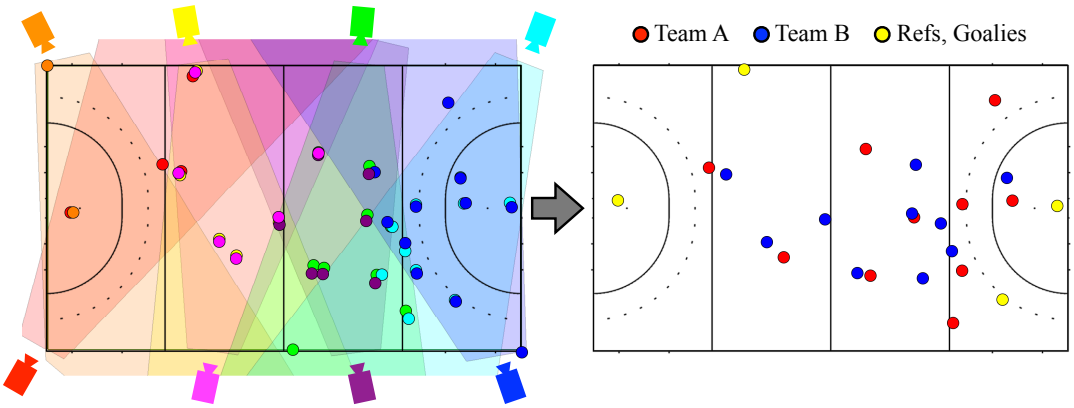


Figure 5.3: Merging the detections from the eight cameras. (Left) The players were detected in each camera using a real-time person detector and their positions were projected into real world co-ordinates. (Right) The detections were then classified into one of the two teams (or discarded if the distance was outside a threshold) and the detections from the cameras were combined to represent the state of the game at each time instant.

colour histograms are sufficient for distinguishing between two teams and are quick to compute and compare, which motivated their use.

Since the camera views overlap in areas, a player may appear in multiple cameras

simultaneously. Therefore, detections corresponding to the same individual from multiple cameras must be identified and combined into a single location when aggregating all the camera detections. The detections from the eight cameras were aggregated by projecting the detected player positions to field co-ordinates using each camera's homography, and merging player detections based on proximity and pixel uncertainty [35]. The Mahalanobis distance between all pairs of detections was computed and any detections which were less than one unit apart that were classified to the same team were clustered into a single detection according to the equations in [28]. The camera layout and merging of detections is illustrated in Fig. 5.3.

The performance of the detector and team classification compared to ground truth annotated frames using precision and recall metrics is shown in Table 5.1. From this table, it can be seen that while recall is high, the team classification has quite low precision in some matches. The poor performance is mainly attributed to non-team-players (referees, goalies, and false-positive player detections caused by background clutter) being misclassified into one of the teams, as they contain a combination of team colours. A more sophisticated representation could be used for modelling the teams as well as non-team image patches, and online learning of the colour models to adapt with changes in illumination would further improve results. From these results, it is evident that the team behaviour representation must be able to deal with a high degree of noise.

5.3 Modelling Team Behaviours

An intuitive representation of team behaviours in sports would be to track all players, maintaining their identity, and the ball. For field-hockey, this would result in a 42 dimensional signal per frame (i.e. 21 objects with x and y coordinates

Match Code	Frames	Precision			Recall		
		Det.	Team A	Team B	Det.	Team A	Team B
10-USA-RSA-1	14352	81.1%	67.2%	77.7%	89.0%	98.3%	98.4%
24-JPN-USA-1	20904	89.5%	91.7%	90.0%	87.5%	95.2%	97.4%
24-JPN-USA-2	7447	85.8%	72.4%	79.7%	90.0%	97.6%	97.0%

Table 5.1: Precision and recall values of the player detector ('Det.') and team classifier separated into 'Team A and 'Team B' (relative to the detected players) after aggregating all cameras.

– 10 field players excluding the goalie $\times$ 2 teams, and the ball). Since the players and ball cannot be reliably tracked over long durations, an alternative is to represent the match via player detections.

Using the proposed player detection and team classification procedure, team behaviours can be modelled as a series of observations, $\mathcal{O}$, where each observation consists of an (x, y) ground location, a timestamp t, and a team affiliation estimate $\tau \in \{\alpha, \beta\}$. At any given time instant t, the set of detected player locations $\mathcal{O}_t = \{x_A, y_A, x_B, y_B, \dots\}$ is of arbitrary length because some players may not have been detected and/or background clutter may have been incorrectly classified as a player. Therefore, the number of player detections at any given frame is generally not equal to the actual number of players, $2P$, where $P = 10$ players per team.

In order to model team behaviours compactly, a method to clean-up noisy detections is needed as well as a representation which exploits the high degree of correlation between players. Player tracks could allow this, but the issue of noisy detections (i.e. missed and false detections) must be overcome.

Tracking players across time is equivalent to generating a vector of ordered player locations $\mathbf{x}_t^\tau = [x_1, y_1, x_2, y_2, \dots, x_P, y_P]^\mathsf{T}$ for each team τ from the noisy detections $\mathcal{O}_t$ at each time instant. The particular ordering of players is arbitrary, but

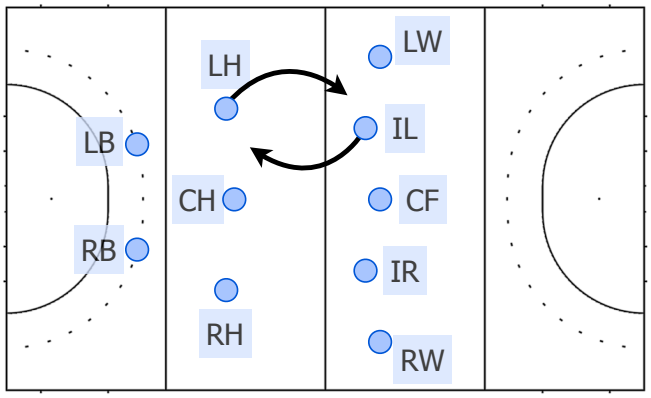


Figure 5.4: The 5:3:2 field-hockey formation. The dynamic nature of the game requires players to switch roles and responsibilities on occasion, for example, the left halfback LH interchanges with the inside left IL to exploit a possible opportunity.

must be consistent across time. It is important to note that $\mathbf{x}_t^\tau$ is not simply a subset of $\mathcal{O}_t$. If a player was not detected, an algorithm must somehow infer the location of the unseen player. To achieve this, the subspace in which player spatial and temporal structure lies must be learnt.

5.3.1 Formations and Roles

As in most team sports, players in a field-hockey team maintain a spatial structure which can be described as a formation. A common field-hockey formation is the 5:3:2 (a line-up weighted towards the offensive, consisting of five offensive players), and defines the set of roles $\mathcal{R} = \{$left back (LB), right back (RB), left halfback (LH), centre halfback (CH), right halfback (RH), inside left (IL), inside right (IR), left wing (LW), centre forward (CF), right wing (RW)$\}$, as illustrated in Figure 5.4. While the team maintains the spatial structure throughout the majority of the match, the dynamic nature of the game involves frequent player role swaps. In Chapter 3, a *role representation* was introduced to handle the role swaps and maintain spatial structure, by assigning roles to players dynamically at

Match Code	Annotated Frames
10-USA-RSA-1	3894
10-USA-RSA-2	8839
24-JPN-USA-1	4855
24-JPN-USA-2	7418
Total	25106

Table 5.2: Details of the manually annotated data. The location, identity and role of each player was manually annotated at each frame for parts of four games from an international field-hockey tournament.

each frame. Mathematically, this is equivalent to permuting the player ordering $\mathbf{x}_t^\tau$, with a A $P \times P$ permutation matrix $\mathbf{P}_t^\tau$ at time t, which describes the players in terms of roles $\mathbf{r}_t^\tau$,

$$\mathbf{r}_t^\tau = \mathbf{P}_t^\tau \mathbf{x}_t^\tau. \tag{5.1}$$

Because the spatial relationships of a formation are defined in terms of roles, and not by players (who frequently swap roles during the game), it is expected that the spatio-temporal patterns in $\{\mathbf{r}_1^\tau, \mathbf{r}_2^\tau, \ldots, \mathbf{r}_T^\tau\}$ would be more compact compared to $\{\mathbf{p}_1^\tau, \mathbf{p}_2^\tau, \ldots, \mathbf{p}_T^\tau\}$. Additionally, it is expected that a team will maintain its formation while moving up and down the field, so the position data expressed relative to the mean (x, y) location of the team should be even more compressible. To test these conjectures, all the players were manually tracked over 25000 time-steps (which equates to $8 \times 25000 = 200,000$ frames across 8 cameras), and a field hockey expert assigned roles to the player locations in each of these frames. A breakdown of the manually labelled data is given in Table 5.2.

Because any observed arrangement of players from team α could also have been observed for players from team β, there is a 180° symmetry in the data. That is, for any given vector of player locations $\mathbf{p}_t^\tau$, there is an equivalent complement $\overleftrightarrow{\mathbf{p}}_t^\tau$ that can be acquired by rotating all (x, y) locations about the centre of the field and swapping the associated team affiliations. By adding the complement of

each formation, the amount of data that can be used for modelling and analysis can be effectively doubled.

For brevity, the analysis is explained in terms of roles $\mathbf{r}_t^\tau$ since the original player ordering $\mathbf{p}_t^\tau$ is just a special non-permuted case $\mathtt{x}_t^\tau = \mathtt{I}$. PCA analysis was conducted on the temporal data series produced by both teams $\{\mathbf{r}_1^\tau, \mathbf{r}_2^\tau, \ldots, \mathbf{r}_{25000}^\tau, \overleftrightarrows^\tau_1, \overleftrightarrows^\tau_2, \ldots, \overleftrightarrows^\tau_{25000}\}$. This was performed to measure how well the low-dimensional representation $\hat{\mathbf{r}}_t^\tau$ matches the original data $\mathbf{r}_t^\tau$ using the L_∞ norm of the residual $\Delta\mathbf{r} = \hat{\mathbf{r}}_t^\tau - \mathbf{r}_t^\tau$:

$$\|\Delta\mathbf{r}\|_\infty = \max(\|\Delta\mathbf{r}(1)\|_2, \ldots, \|\Delta\mathbf{r}(P)\|_2) \tag{5.2}$$

where $\|\Delta\mathbf{r}(p)\|_2$ is the L_2 norm of the p^{th} x and y components of $\Delta\mathbf{r}$. The L_∞ norm was chosen instead of the L_2 norm because large deviations may signify very different formations, e.g. a single player could be breaking away to score. Figure 5.5 illustrates how both $\mathbf{p}_t^\tau$ and $\mathbf{r}_t^\tau$ are quite compressible on the training data. When testing on unseen data (with role labels), the dynamic role-based ordering $\mathbf{r}_t^\tau$ is much more compressible than the static ordering $\mathbf{p}_t^\tau$. Relative positions are more compressible than absolute positions in both orderings.

In Figure 5.6, the mean formations for the identity and role representation are shown. It can be seen that the role representation has a more uniform spread between the players, while the identity representation has a more crowded shape, which highlights the constant swapping of roles during a match.

5.3.2 Incorporating Adversarial Behaviour

A player's movements are correlated not only to teammates but to opposition players as well. Therefore, the player location data can be further compressed if the locations of players on teams A and B are concatenated into a single vector

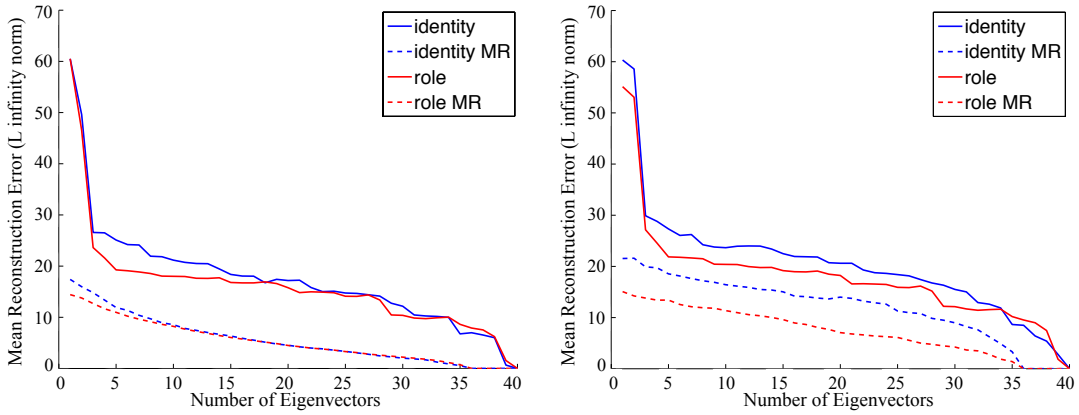


Figure 5.5: Plots showing the reconstruction error as a function of the number of eigenvectors used to reconstruct the signal using the L_∞ norm for original and mean-removed (MR) features for both identity and role representations on training data (left) and unseen test data (right).

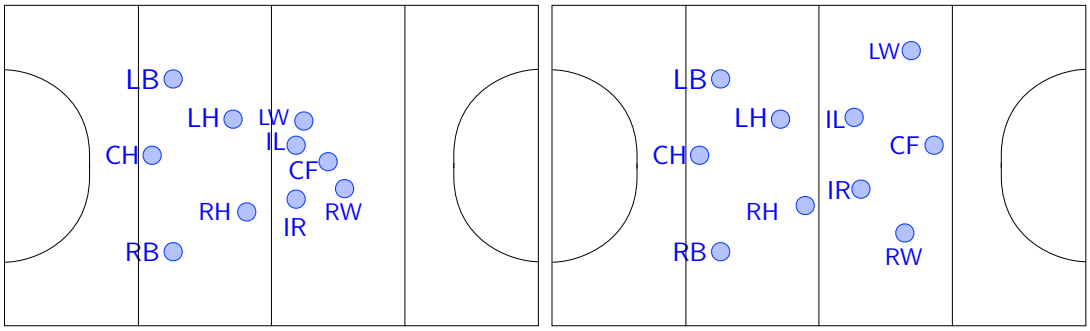


Figure 5.6: Examples showing the difference between the mean formations using the: (left) identity and (right) role representations on one of the matches.

$\mathbf{r}_t^{AB} = [\mathbf{r}_t^A, \mathbf{r}_t^B]^\mathsf{T}$.

In terms of compressibility, Table 5.3 shows that using an adversarial representation, incorporating both teams, gains better compressibility for identity and role representations, and that using both a role and adversarial representation yields the most compressibility.

	Features required to represent 95% of the original signal	
Representation	**Identity**	**Role**
Single Team	30%	25%
Adversarial Teams	20%	**15%**

Table 5.3: The compressibility of different representations, in terms of the percentage of features required to represent 95% of the original signal when using PCA.

5.4 Cleaning-Up Noisy Data

5.4.1 Spatio-temporal Bilinear Basis Model

The representation of time-varying spatial data is a well-studied problem in computer vision (see [25] for overview). Recently, Akhter et al. [3] presented a bilinear spatio-temporal basis model which captures and exploits the dependencies across both the spatial and temporal dimensions in an efficient manner. Such a model can be applied to represent player tracking data and provide a lower dimensionality signal but requires appropriate models of the spatial and temporal basis specific to groups which are presented in this chapter.

Given P players per team, the role-based adversarial representation, $\mathbf{r}$, can be represented as a spatio-temporal structure $\mathbf{S}$ with $2P$ total players sampled at F time instances:

$$\mathbf{S}_{F\times 2P} = \begin{bmatrix} x_1^1 & \dots & x_{2P}^1 \\ \vdots & & \vdots \\ x_1^F & \dots & x_{2P}^F \end{bmatrix} \tag{5.3}$$

where x_j^i denotes the jth index within the role representation at the ith time instant. Thus, the time-varying structure matrix $\mathbf{S}$ contains $2FP$ parameters.

This representation of the structure is an over parametrisation because it does not take into account the high degree of regularity generally exhibited by motion data. One way to exploit the regularity in spatio-temporal data is to represent the 2D formation or shape at each time instance as a linear combination of a small number of shape basis vectors $\mathbf{b}_j$ weighted by coefficients ω_j^i as $\mathbf{s}^i = \sum_j \omega_j^i \mathbf{b}_j^T$ [22, 33]. An alternative representation of the time-varying structure is to model it in the trajectory subspace, as a linear combination of trajectory basis vectors, $\boldsymbol{\theta}_i$ as $\mathbf{s}_j = \sum_i a_i^j \boldsymbol{\theta}_i$, where a_i^j is the coefficient weighting each trajectory basis vector [1, 134]. As a result, the structure matrix can be represented as either :

$$\mathbf{S} = \boldsymbol{\Omega}\mathbf{B}^T \quad \text{or} \quad \mathbf{S} = \boldsymbol{\Theta}\mathbf{A}^T, \tag{5.4}$$

where $\mathbf{B}$ is a $P \times K_s$ matrix containing K_s shape basis vectors, each representing a 2D structure of length $2P$, and $\boldsymbol{\Omega}$, is an $F \times K_s$ matrix containing the corresponding shape coefficients ω_j^i; and $\boldsymbol{\Theta}$ is an $F \times K_t$ matrix containing K_t trajectory basis as its columns, and $\mathbf{A}$ is a $2P \times K_t$ matrix of trajectory coefficients. The number of shape basis vectors used to represent a particular instance of motion data is $K_s \leq \min\{F, 2P\}$, and $K_t \leq \{F, 2P\}$ is the number of trajectory basis vectors spanning the trajectory subspace.

Both representations of $\mathbf{S}$ are over parametrisations because they do not capitalise on either the spatial or temporal regularity. As $\mathbf{S}$ can be expressed exactly as $\mathbf{S} = \boldsymbol{\Omega}\mathbf{B}^T$ and also $\mathbf{S} = \boldsymbol{\Theta}\mathbf{A}^T$, then there exists a factorisation:

$$\mathbf{S} = \boldsymbol{\Theta}\mathbf{C}\mathbf{B}^T \tag{5.5}$$

where $\mathbf{C} = \boldsymbol{\Theta}^T\boldsymbol{\Omega} = \mathbf{A}^T\mathbf{B}$ is a $K_t \times K_s$ matrix of spatio-temporal coefficients. This equation describes the bilinear spatio-temporal basis, which contains both shape and trajectory bases linked together by a common set of coefficients.

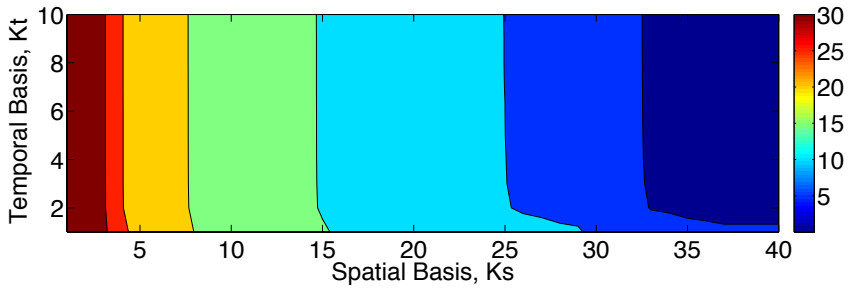


Figure 5.7: Plot showing the mean reconstruction error on the test data as the number of temporal basis (K_t) and spatial basis (K_s) vary for 5 second plays (i.e. $K_{t\max} = 150$). The plot is magnified to show the first 10 temporal basis, highlighting that only $K_t = 3$ is required to represent coarse player motion.

Due to the high degree of temporal smoothness in the motion of humans and sports players, a predefined analytical trajectory basis can be used without significant loss in representation. A particularly suitable choice of a conditioning trajectory basis is the Discrete Cosine Transform (DCT) basis, which has been found to be close to the optimal Principal Component Analysis (PCA) basis if the data is generated from a stationary first-order Markov process [116]. Given the high temporal regularity present in almost all human motion, it has been found that the DCT is an excellent basis for trajectories of faces [2, 3] and bodies [7]. Figure 5.7 shows that due to the highly structured nature of the game, and the fact that human motion over short periods of time is very simple, an enormous reduction in dimensionality can be achieved, especially in the temporal domain. From this, a 5 second play can be effectively represented using $K_t = 3$ and $K_s = 33$ with a maximum error of less than 2 meters. In terms of dimensionality reduction, this means temporal signals can be represented using $3 \times 33 = 99$ coefficients. For 5 second plays, this means a reduction of over 60 times. Even greater compressibility can be achieved on longer plays.

5.4.2 The Assignment Problem

In the previous section, manually annotated roles were used for analysis. Now, the problem of automatically assigning roles to an arbitrary ordering of player locations $\mathbf{p}_t^\tau$ is outlined. Assuming a suitably similar vector $\hat{\mathbf{r}}^\tau$ of player locations in role order exists, the optimal assignment of roles is defined by finding the permutation matrix $\mathtt{x}_t^{\tau\star}$ which minimises the square L_2 reconstruction error:

$$\mathtt{x}_t^{\tau\star} = \arg\min_{\mathtt{x}_t^\tau} \|\hat{\mathbf{r}}^\tau - \mathtt{x}_t^\tau \mathbf{p}_t^\tau\|_2^2. \tag{5.6}$$

This is the linear assignment problem where an entry $\mathtt{C}(i,j)$ in the cost matrix is the Euclidean distance between role locations

$$\mathtt{C}(i,j) = \|\hat{\mathbf{r}}^\tau(i) - \mathbf{p}_t^\tau(j)\|_2. \tag{5.7}$$

The optimal permutation matrix can be found in polynomial time using the Hungarian (or Kuhn-Munkres) algorithm [79].

5.4.3 Assignment Initialisation

To solve the assignment problem, a reference formation is needed to compare to. Using the mean formation (see Figure 5.6) is a reasonable initialisation as the

Representation	**Prototype**	**Hit Rate**	
		Team A	Team B
Identity	Mean Formation	38.36	29.74
	Codebook	49.10	37.15
Role	Mean Formation	49.47	50.30
	Codebook	**74.18**	**69.70**

Table 5.4: Accuracy of the assignment using a mean formation versus using a codebook of formations.

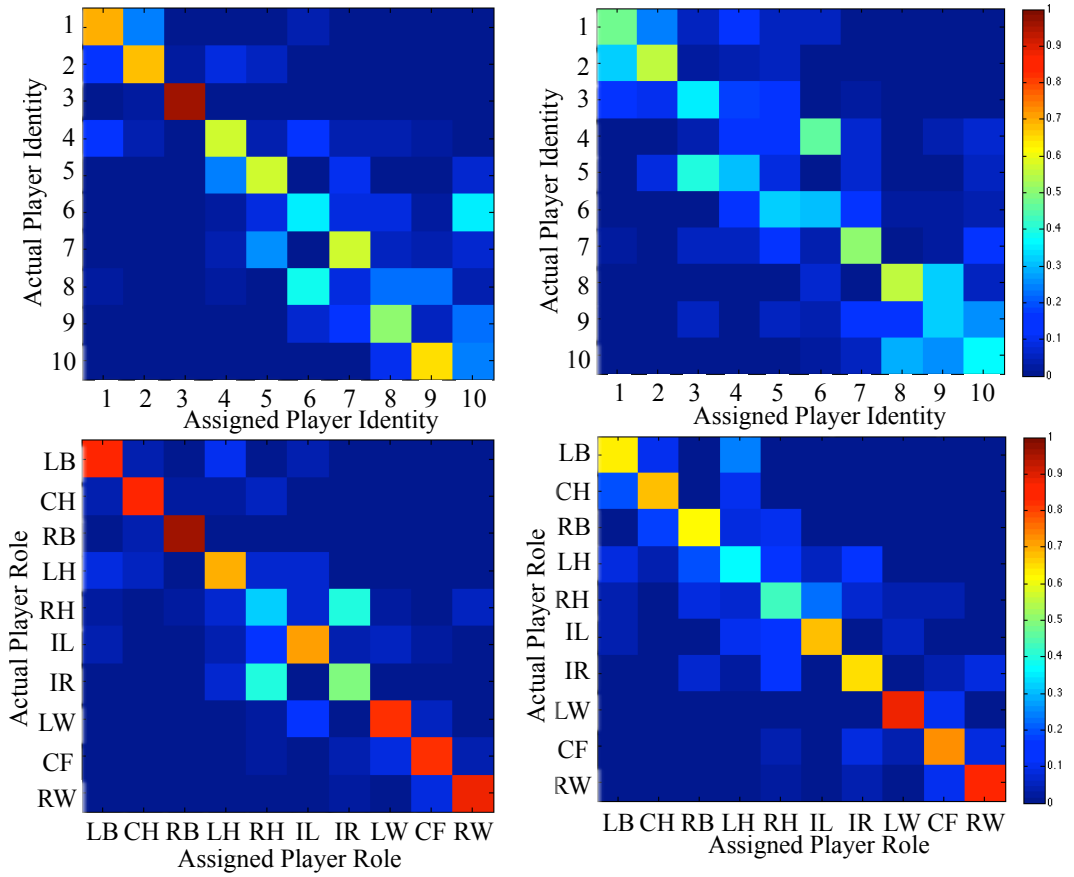


Figure 5.8: Confusion matrices showing the hit-rates for correctly assigning identity (top row) and role (bottom) for Team1 (left) and Team2 (right) on the test set.

team should maintain that basic formation in most circumstances. However, in different areas of the field there are subtle changes in formation due to what the opposition is doing as well as the game-state. To incorporate these semantics, a codebook of formations was used which consists of every formation within the training set. This mapping is difficult to find as the input features have no role assignment, and every permutation would have to be tested to find the matching template, which is highly inefficient. Using the assignment labels of the training data, a mapping matrix $\mathbf{W}$ can be learnt from the mean and covariances of the training data to a labelled formation via the linear transform $\mathbf{X} = \mathbf{W}^T\mathbf{Z}$. Given N training examples, $\mathbf{W}$ can be learnt by concatenating the mean and covariance

into an input vector $\mathbf{z}_n$, which corresponds to the labelled formation $\mathbf{x}_n$. All these features are compiled into the matrices $\mathbf{X}$ and $\mathbf{Z}$, and then linear regression is used to find $\mathbf{W}$ by solving:

$$\mathbf{W} = \mathbf{X}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \lambda\mathbf{I})^{-1} \tag{5.8}$$

where λ is the regularisation term. Using this approach, a labelled formation can be estimated from the training set which best describes the current unlabelled one. In terms of assignment performance on the test set, this approach works very well compared to using the mean formation for both the identity and role labels as can be seen in Table 5.4. The confusion matrices for both Team A and Team B for both representations are shown in Figure 5.8. It is worth noting that the role representation gave far better results than the identity representation, which is not surprising seeing that only spatial location is used. In terms of the role representation (bottom two plots), it can be seen that there is little confusion between the 3 defenders (LB, CH, RB) and the 3 forwards (LW, CF, RW). The midfield 4 (LH, RH, IL, IR) tend to interchange position a lot, causing high confusion. Noticeably, there is a discrepancy between Team A and Team B which is understandable in this case as Team B interchanges positions more than twice the amount than Team A does upon analysis of the ground-truth.

5.5 Interpreting Noisy Data

In practice, perfect data from a vision-system can not be obtained so this method has to be robust against both missed and false detections. Given the four annotated matches (presented in Table 5.2), the precision and recall rates for the detector and the team affiliation are given in the left side of Table 5.5. In this work, a detection was considered to be correct if a player was within two meters

	Raw Detections		With Assignment	
	Precision	Recall	Precision	Recall
Detections	77.49	89.86	**91.90**	80.46
Team A	72.54	86.14	**86.69**	74.17
Team B	79.84	89.66	**92.91**	82.85

Table 5.5: Precision-Recall rates for the raw detections (left) and with the initialised assignments (right).

of a ground-truth label.

5.5.1 Assigning Noisy Detections

Assuming that the majority of player detections and team affiliations are correct, the task is to assign role labels to the detections and discard any detections deemed to be too noisy. To determine whether or not an assignment should be made or if the detection should be discarded, some type of game context feature can be used, such as which part of the field the players are located. To do this, a similar strategy to the one proposed in Section 5.4.3 was employed. Instead of learning the mapping from the clean features $\mathbf{Z}$, the mapping is learnt from the noisy features $\mathbf{Z}_{\text{noisy}}$. Because the player detector has systematic errors (there are some "black-spots" on the field due to reduced camera coverage, and game situations where players bunch together), the number of players detected from the system was incorporated in addition to the mean and covariance in the noisy game context feature $\mathbf{z}_{\text{noisy}}$. This is then used to learn $\mathbf{W}_{\text{noisy}}$. This is possible because of the assumption that the clean centroid is a good approximation to the noisy centroid which is shown to be a valid assumption in Figure 5.9. Using this assumption, a reasonable prototypical formation can be obtained to make the role assignments.

Using the estimated prototype formation, the assignment is then performed using

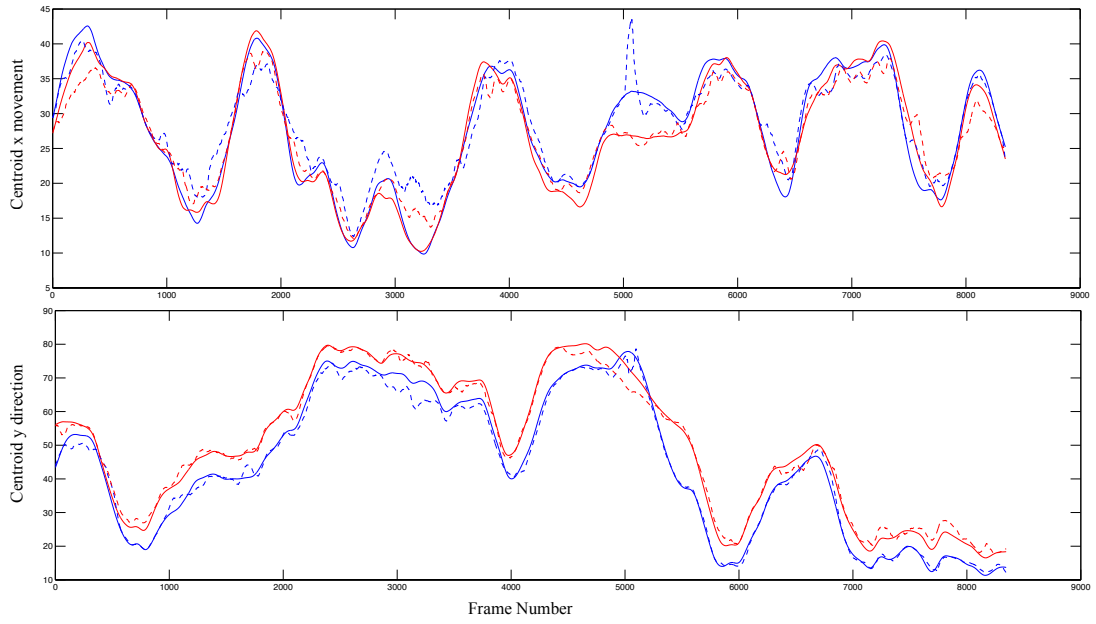


Figure 5.9: As the centroids of both the clean (solid) and noisy (dashed) of both teams (blue = Team1, red = Team2) are roughly equivalent, a mapping matrix is learnt using linear regression to find a formation from the training set which can best describe the noisy test formation.

the Hungarian algorithm. This is challenging however, as there may be missed or false detections which alters the one-to-one mapping between the prototype and input detections. To counter this, an "exhaustive" approach was employed, where if there are fewer detections than the number of players in the prototype, all the possible combinations that the labels could be assigned are found, and the combination which yielded the lowest cost from the assignments is made. Conversely, if there are more detections than the number of players, all the possible combinations that the detections could be are found and then the combination of detections which had the lowest cost is used.

Sometimes there may be false positives which means that even though there are 10 detections for a team only 7 or 8 may be valid candidates. Employing the exhaustive approach greatly improves the precision rate, while the recall rate decreases which is to be expected (see right side of Table 5.5). Even despite the

drop in recall, roles are assigned reasonably well (over 55% compared to 66% on the clean data) as can be seen in Table 5.6.

5.5.2 De-noising the Detections

While the precision and recall rates from the detector are relatively high, to do useful analysis such as formation and play analysis, a continuous estimate of the player label at each time step is necessary. This requires a method which can de-noise the signal - i.e. a method which can impute missing data and filter out false detections. Given the spatial basis, the bilinear coefficients and an initial estimate of the player labels, an Expectation Maximisation (EM) algorithm can be used to de-noise the detections. The employed approach is similar to that proposed by Akhter et al. [3]. Using this approach, the expectation step is simplified to making an initial hard assignment of the role labels which can be achieved using the method described in the previous section. From this initialisation, an initial guess of $\hat{\mathbf{S}}$ is acquired. In the maximisation step, the coefficients are calculated using $\mathbf{C} = \Theta^T \hat{\mathbf{S}} \mathbf{B}$, and then $\mathbf{S}$ is estimated from the new $\mathbf{C}$ as well as the spatial and temporal bases $\mathbf{B}$ and Θ. Examples of the cleaned up detections using this approach are shown in Figure 5.10.

Using this procedure provides an estimate of the continuous trajectories for all 10 roles effectively. As the recall rate of the de-noised data is 100%, the proposed

	Correct	Incorrect	Missed	Hit Rate
Team A	41.89	32.89	25.22	56.02
Team B	45.92	35.56	18.53	56.36

Table 5.6: The compressibility of different representations. The column on the far right gives the effective hit-rate (i.e. missed detections omitted) of the correct assignments.

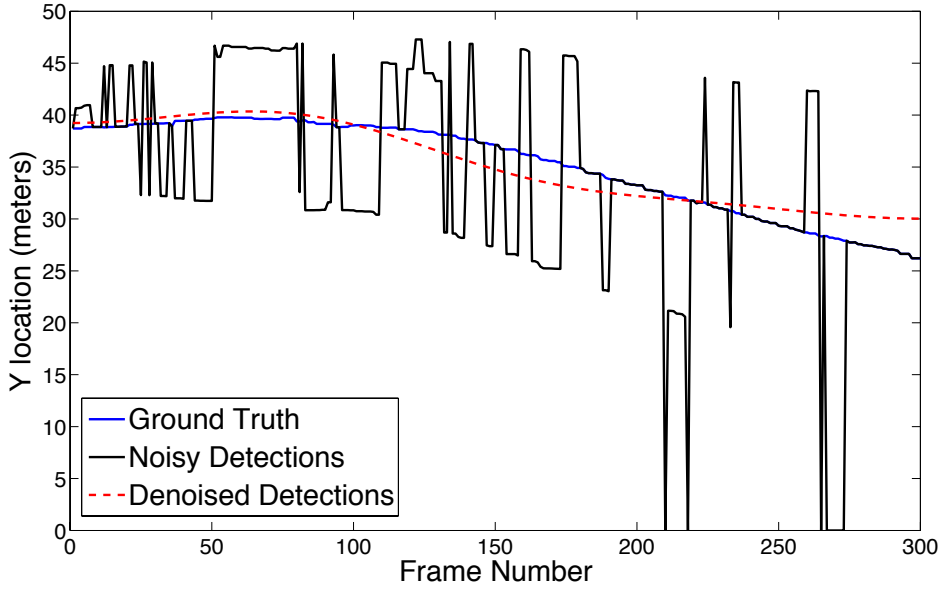


Figure 5.10: Given the noisy detections (black), the bilinear model can be used to estimate the trajectory of each player over time. It can be seen that the estimate (red) is close to the ground-truth (blue).

method is evaluated in terms of how precise this method is at inferring player position based on their label. To test this, the precision rate for the detections and the de-noised detections was calculated against a distance threshold (i.e. the minimum distance a player had to be to ground-truth to be recognised as a correct detection). The results are shown in Figure 5.11. As can be seen from these figures, the detections from the player detector are very accurate and do not vary with respect to the error threshold (i.e. it either detects a player very precisely or not at all). Conversely, the de-noised data is heavily smoothed due to the bilinear model, so some of the finer details are lost to gain a continuous signal.

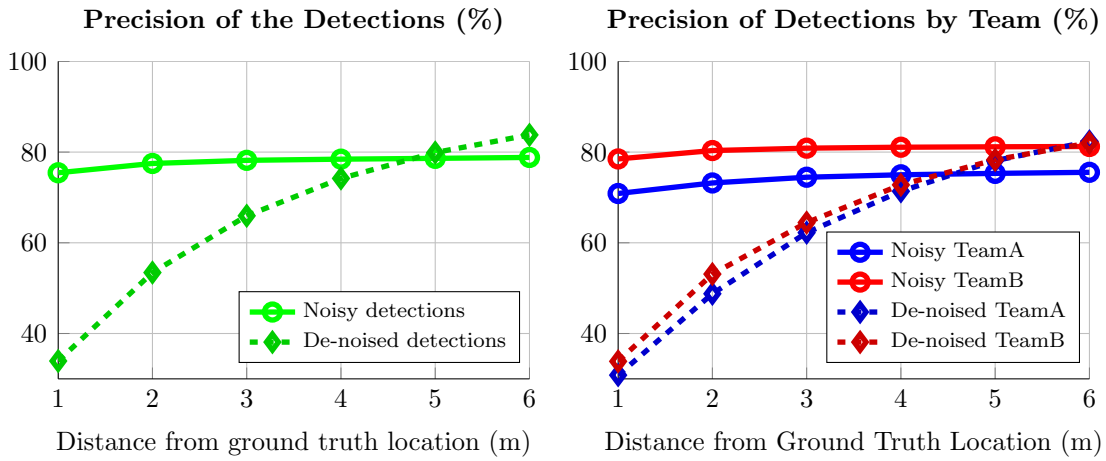


Figure 5.11: Precision accuracy vs the distance threshold from ground-truth for: (left) the overall detections, (right) the detections based on team affiliation. The solid lines refer to the raw, noisy detections and the dashed lines refer to the de-noised signal.

5.5.3 Formation and Play Analysis

To demonstrate the usefulness of the cleaned-up signal, cluster analysis was conducted on both static formations and dynamic plays to see whether results achieved with manually labelled roles could be replicated. The first analysis conducted was to find the top three formations that could best describe the test data (i.e. the 3 most likely formations that occurred). The results are shown in Figure 5.12. From the figure it can be seen that despite small differences, the results are similar to what is obtained from manually labelled data with the first formation being identical and formations 3 and 2 are reversed. A similar trend was observed for the play analysis where 10 second plays were clustered (see Figure 5.13). As can be seen from the de-noised data, the bilinear model has smoothed out the trajectory, although it is unrealistic in some cases. Despite this, similar results were obtained, and the analysis can be done with a fraction of the amount of features due to the high compressibility of the signal (D = 200 vs D = 12000).

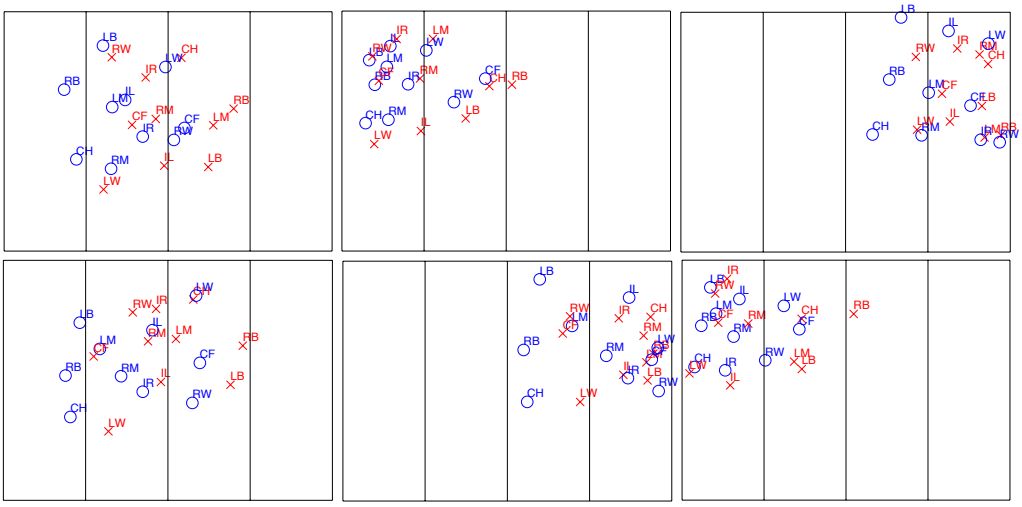


Figure 5.12: Cluster analysis of the top three formations (1-3, ordered left-to-right) which best represent the test data using manually labelled data (top) and the de-noised data (bottom). The blue team is attacking from left-to-right.

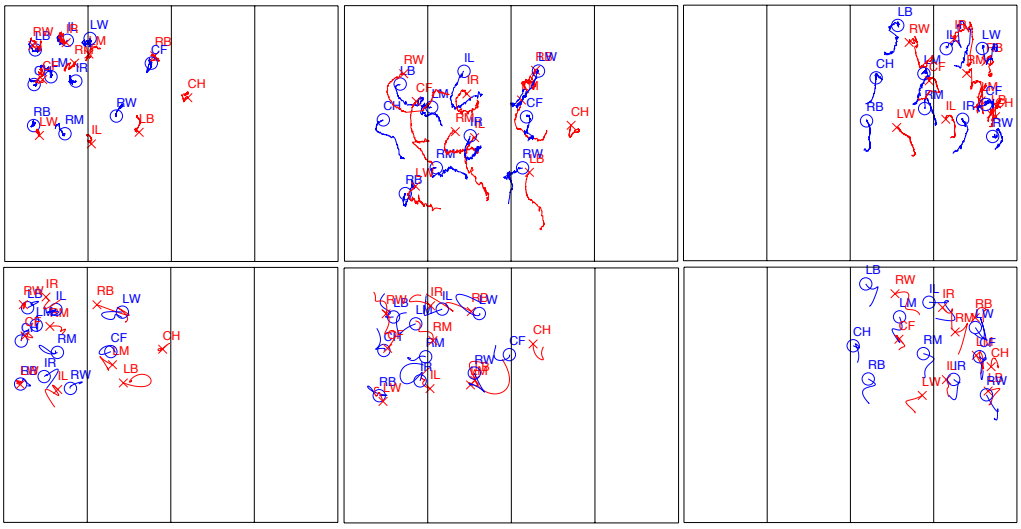


Figure 5.13: Cluster analysis of the top 10-second plays on the test data using manually labelled data (top) and our de-noised data (bottom). The x's and the o's refer to the position of the player at the end of the 10 second play. The blue team is attacking from left-to-right.

5.6 Summary

Accurately tracking objects over long durations of time is an unsolved computer vision problem, and prevents automated analysis of groups using traditional representations based on tracks. In this chapter, instead of using manually labelled data which is very time consuming to acquire, group behaviours were modelled directly from raw detections. A representation which utilised role labels to exploit the heavy spatio-temporal correlations that exist within adversarial domains was presented. As the representation is highly correlated in both space and time, it was shown that a spatio-temporal bilinear basis model could leverage this trait to compress the incoming signal by up to two orders of magnitude without much loss of information. This makes analysis and clustering tasks tractable on temporal group behaviour data. The approach was evaluated on approximately 200,000 frames of field-hockey data from a state-of-the-art real-time player detector and results were presented for clustering formations and plays, achieving similar results to manually cleaned-up data in addition to having a 60 times reduction in dimensionality. Following the demonstration of compressibility of adversarial spatio-temporal data, the use of the bilinear model was shown to effectively clean up noisy player detections, which enabled analysis of static formations as well as temporal plays. The proposed representation can be applied to other spatio-temporal data with repetitive patterns (particularly adversarial multi-agent data), however the chosen spatial and temporal basis would depend on the structure and motion patterns in the data.

Chapter 6

Recognising Team Activities from Noisy Data

6.1 Introduction

Recently, vision-based systems have been deployed in professional sports to track the ball and players to enhance analysis of matches. Due to their unobtrusive nature, vision-based approaches are preferred to wearable sensors (e.g. GPS or RFID sensors) as they do not require players or balls to be instrumented prior to matches. Unfortunately, in continuous team sports where players need to be tracked continuously over long-periods of time (e.g. 35 minutes in field-hockey or 45 minutes in soccer), current vision-based tracking approaches are not reliable enough to provide fully automatic solutions. As such, human intervention is required to fix-up missed or false detections. In instances where a human can not intervene due to the sheer amount of data being generated, this data can not be used due to the missing/noisy data. In this chapter, two representations based on raw player detections (and not tracking) are investigated and shown to be robust

to missed and false detections. Specifically, it is shown that both team occupancy maps and centroids can be used to detect team activities, while occupancy maps can be used to retrieve specific team activities. An evaluation on over 8 hours of field hockey data captured at a recent international tournament demonstrates the validity of the proposed approach.

6.2 Related work

Due to the host of military, surveillance and sport applications, research into recognising group behaviour has recently increased dramatically. Outside of the sports realm, most of this work has focussed on dynamic teams (i.e. where individual agents can leave and join teams over the period of the observations). An initial approach was to recognise the activities of individual agents and then combine these to infer group activities [13]. Sukthankar and Sycara recognised group activities as a whole but pruned the size of possible activities by using temporal ordering constraints and agent resource dependencies [127, 128]. Sadilek and Kautz [122] used GPS locations of multiple agents in a "capture the flag" game to recognise low-level activities such as approaching and being at the same location. All of these works assume that the position and movements of all agents are known, and that all behaviours can be mapped to an activity within the library. Recently, Zhang et al. [145] used a "bag of words" and Support Vector Machine (SVM) approach to recognise group activities on the Mock Prison dataset [30].

Sport related research mostly centres on low-level activity detection with the majority conducted on American Football. In the seminal work by Intille and Bobick [69], they recognised a single football play *pCurl51*, using a Bayesian network to model the interactions between the players trajectories. Li et al. [86],

modelled and classified five offensive football plays (dropback, combo dropback, middle run, left run, right run). Siddiquie et al. [124], performed automated experiments to classify seven offensive football plays using a shape (HoG) and motion (HoF) based spatio-temporal features. Instead of recognising football plays, Li and Chellapa [85] used a spatio-temporal driving force model to segment the two groups/teams using their trajectories. Researchers at Oregon State University have also done substantial research in the football space [64, 65, 126] with the goal of automatically detecting offensive plays from a raw video source and transferring this knowledge to a simulator. For soccer, Kim et al. [75] used the global motion of all players in a soccer match to predict where the play will evolve in the short-term. Beetz et al. [17] developed the *automated sport game models* (ASPOGAMO) system which can automatically track player and ball positions via a vision system. Using soccer as an example, the system was used to create a heat-map of player positions (i.e. which area of the field did a player mostly spend time in) and also has the capability of clustering passes into low-level classes (i.e. long, short etc.), although no thorough analysis was conducted due to a lack of data. In basketball, Perse et al. [110] used trajectories of player movement to recognise three type of team offensive patterns. Morariu and Davis [102] integrated interval-based temporal reasoning with probabilistic logical inference to recognise events in one-on-one basketball. Hervieu et al. [63] also used player trajectories to recognise low-level team activities using a hierarchical parallel semi-Markov model.

An enormous amount of research interest has used broadcast sports footage for video summarisation in addition to action, activity and highlight detection [17, 43, 58, 68, 80, 92, 101, 141], but given that these approaches are not automatic (i.e. the broadcast footage is generated by humans) and that the telecasted view captures only a portion of the field, analysing groups has been impossible because some individuals are normally out of frame. Although similar in spirit

to the research mentioned above, the work presented in this chapter differs as: 1) it relies only on player detections rather than tracking, and 2) it is evaluated across many matches (7 compared to 1).

6.3 Detection Data

6.3.1 Field-Hockey Test-Bed

To enable this research, player detection data was captured from the field-hockey test bed described in Section 5.2.1 and seven complete field-hockey matches were analysed. In this test-bed, each of the 8 cameras is connected to a computer which processes that camera's video footage and is used to detect the player positions and their team from that camera. This is then relayed to a central hub via optic fibre where the detections are merged, and can be analysed online for tasks such as activity recognition. As the player detector has a latency of only 1 frame, analysis can be performed in real-time. Over seven complete field-hockey matches were collected and analysed from a recent field hockey tournament (consisting of over 8 hours of match data for each of the 8 HD cameras). The analysed matches are listed in Table 6.1, along with the number of frames annotated with players' team and field position (x,y). Due to the enormous amount of time it takes to manually label player tracks during a match, the labelling effort was limited to four halves from three matches. Team activities were labelled for seven complete matches as can be seen in this table.

The procedure for detecting players and assigning team affiliation was detailed in Section 5.2.2, and the precision and recall values of the detector and team affiliation of the data were presented in Table 5.1. It is evident that while the detector has high recall rates of the player positions (ranging from 87.5% to

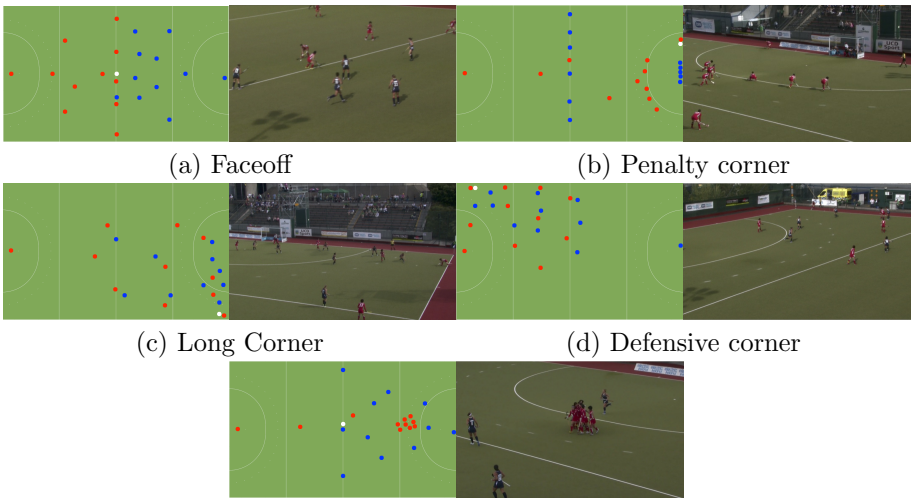
		Frames Annotated	
Match code	Activities	1st Half	2nd Half
1-JPN-USA	✓	-	-
2-RSA-SCO	✓	-	-
5-USA-SCO	✓	-	-
9-JPN-SCO	✓	-	-
10-USA-RSA	✓	14352	-
22-RSA-IRL	-	17861	-
23-ESP-SCO	✓	-	-
24-JPN-USA	✓	20904	7447

Table 6.1: Itemised list of analysed field-hockey data. Frames annotated refers to the number of frames where players' location and team identity were manually annotated for quantifying the detector's accuracy.

90.0%), the team classification has quite low precision in some matches (ranging from 67.2% to 91.7%). The poor performance is mainly due to false positive detections being misclassified into one of the teams. From this, it is evident that the team behaviour representation must be able to deal with a high degree of noise.

6.3.2 Team Activity Labels

In this chapter, classification and retrieval of common activities that occur in field-hockey is performed from automatically extracted detection data. Seven complete matches were annotated with these activities, which are listed along with their frequency count in Table 6.2 for each match half. Pictorial and broadcast examples of the five activities are shown in Figure 6.1. Each of these five activities correspond to activities or statistics that an analyst would label during a game. As each activity has distinctive spatial locations and motion patterns, the reliability of labelling these activities is very high. In this work, the aim is to automatically detect these activities based solely on noisy player detections (i.e.



(a) Faceoff (b) Penalty corner

(c) Long Corner (d) Defensive corner

(e) Goal

Figure 6.1: Diagrams and examples of structured plays that occur in field-hockey

there is no ball or player identity information).

	Face off	Pen. Cnr	Goal	Long Cnr		Def Cnr	
				(L)	(R)	(L)	(R)
1-JPN-USA-1	3	2	2	11	5	4	4
1-JPN-USA-2	2	6	1	4	10	7	3
2-RSA-SCO-1	2	4	2	11	4	3	3
2-RSA-SCO-2	3	9	2	3	12	4	3
5-USA-SCO-1	3	4	2	7	4	1	7
5-USA-SCO-2	3	8	2	3	3	2	2
9-JPN-SCO-1	2	4	2	8	7	5	2
9-JPN-SCO-2	1	1	0	10	10	6	0
10-USA-RSA-1	5	9	5	5	5	8	0
10-USA-RSA-2	6	4	5	6	7	4	1
23-ESP-SCO-1	3	4	2	7	6	1	1
23-ESP-SCO-2	3	7	2	9	5	2	1
24-JPN-USA-1	4	3	3	9	6	5	1
24-JPN-USA-2	2	2	1	5	9	7	6
Total	42	67	31	98	93	59	34

Table 6.2: Activity frequency in each match half

6.4 Representing Team Behaviours

Team sports like field-hockey are played over a very large spatial area. An intuitive representation would be to track all players (maintaining their identity) and the ball, which would result in a 46 dimensional signal (i.e. 23 objects in x and y coordinates – 11×2 players, 1 ball). Since it is not possible to reliably and accurately track the player and ball over long durations (e.g. 35 mins), an alternative is to represent the match via player detections. By using detections, the issue of tracking is overcome, but the player identity component of the signal is removed as a consequence so another method is needed to maintain feature correspondences. In this section two representations are proposed to handle this: occupancy maps and centroid descriptors.

6.4.1 Team Occupancy Maps

The team occupancy map descriptor, $\mathbf{x}_t^o$, is a quantised occupancy map of the player positions on the field for each team represented at time t. Given the locations of the players from the player detector system and assigned team affiliation labels, an occupancy map can be constructed for each frame by splitting the 91.4 m × 55.0 m field into K spatial bins, and counting how many player detections for that team occupy each location. The dimensionality of the formation descriptor is equal to twice the number of bins (i.e. $K \times 2$) so that both teams A and B are accounted for, resulting in $\mathbf{x}_i^o = [a_1, \ldots, a_K; b_1, \ldots, b_K]$, where a_k and b_k are the player counts in bin k for teams A and B respectively. These occupancy maps can then be used to represent team activities by concatenating the vectors from each frame.

Depending on the level of complexity of the activity that must be recognised,

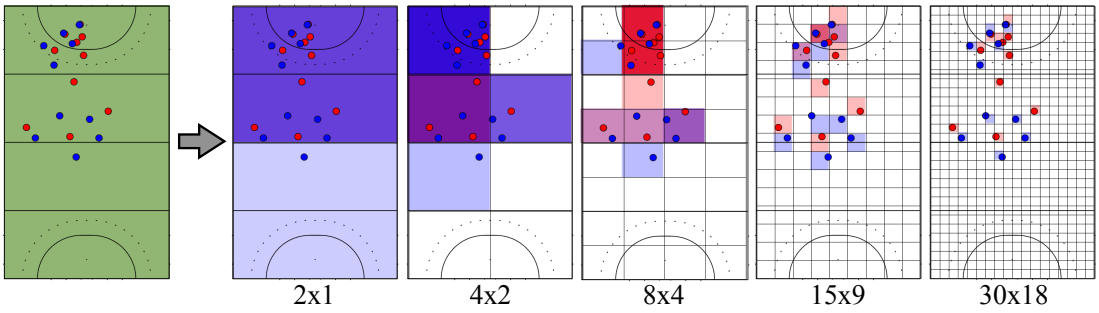


Figure 6.2: Example team occupancy maps for different descriptor sizes.

varying descriptor sizes (coarse/fine) can be used. Five different descriptor sizes were evaluated: $K = 2$ (2×1), $K = 8$ (4×2), $K = 32$ (8×4), $K = 135$ (15×9) and $K = 540$ (30×18), with examples illustrated in Fig. 6.2. These values were chosen to approximately split the grid into square regions, and provide a range of descriptor sizes representing varying levels of quantisation. The different quantisations represent how much tolerance there is in player's positions (e.g. in 15×9 quantisation, each occupancy represents an area of approximately 6 m^2). Since an activity can occur for either team, the occupancy maps are compared in both orientations, ($\mathbf{x}^o = [\mathbf{a}, \mathbf{b}]^T$, and $\mathbf{x}^o = [\mathbf{b}_{rot}, \mathbf{a}_{rot}]^T$, where $\mathbf{a}_{rot}$ represents a rotation of the field by 180° for team a's formation descriptor, so that the new descriptor is given by $\mathbf{a}_{rot}[k] = \mathbf{a}[K + 1 - k]$, for $k = 1, 2, \ldots, K$).

6.4.2 Team Centroid Representation

Given the player detections and their team affiliations, the centroid representation, $\mathbf{x}_i^c$ is found by calculating the mean and covariance of the player positions for each team. As with the team occupancy representation, the centroid features are compared in both orientations, and the rotated positions are given by $x_{rot} = 91.4 - x$ and $y_{rot} = 55.0 - y$ (where 91.4 m $\times$ 55.0 m are the dimensions of the field and x and y are the positions on the field). An example of the team centroid representation is depicted in Figure 6.3.

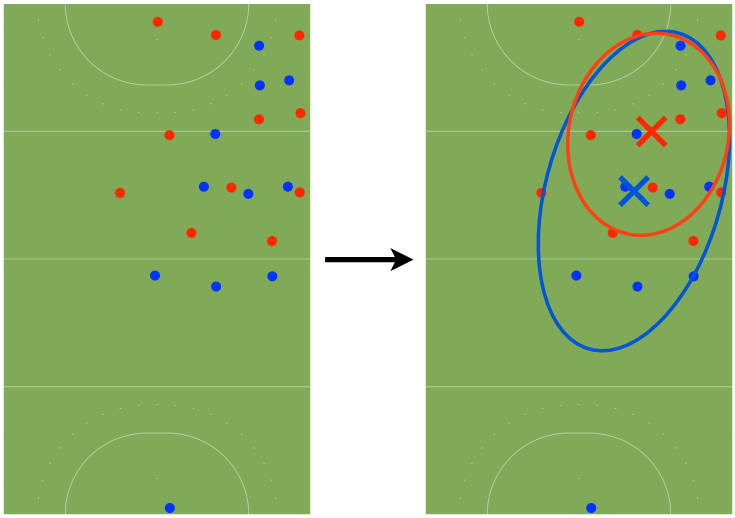


Figure 6.3: Team centroid representation overlaid on the player detections. The 'x' represents the mean position for each team (i.e. the centroids), and the ellipses represent the covariances of their positions.

6.5 Recognising Team Activities

6.5.1 Isolated Activity Recognition

To evaluate the different representations, a series of isolated activity recognition experiments were conducted. The occupancy map and centroid representations were used to recognise five activities, corresponding to important game states in field-hockey, shown in Figure 6.1. As these activities coincide with a single event (e.g. the ball crossing the out line, or a goal being scored), they do not have distinct onset and offset times. To account for this, 10 second play clips were extracted with the labelled event taken as the start of the activity. Since an activity can occur for either team, the template descriptors were compared to the activity template in both orientations, and the minimum distance value was taken as the distance measure.

Seven full matches (corresponding to over 8 hours of game play), were annotated

	Face Off	Penalty Corner	Goal	Long Corner		Defensive Corner	
				(L)	(R)	(L)	(R)
1-JPN-USA-1	3	2	2	11	5	4	4
1-JPN-USA-2	2	6	1	4	10	7	3
2-RSA-SCO-1	2	4	2	11	4	3	3
2-RSA-SCO-2	3	9	2	3	12	4	3
5-USA-SCO-1	3	4	2	7	4	1	7
5-USA-SCO-2	3	8	2	3	3	2	2
9-JPN-SCO-1	2	4	2	8	7	5	2
9-JPN-SCO-2	1	1	0	10	10	6	0
10-USA-RSA-1	5	9	5	5	5	8	0
10-USA-RSA-2	6	4	5	6	7	4	1
23-ESP-SCO-1	3	4	2	7	6	1	1
23-ESP-SCO-2	3	7	2	9	5	2	1
24-JPN-USA-1	4	3	3	9	6	5	1
24-JPN-USA-2	2	2	1	5	9	7	6
Total	42	67	31	98	93	59	34

Table 6.3: Frequency of the annotated activities in each match half.

with the 5 activities of interest: face-offs, penalty corners, goals, long corners and defensive corners as shown in Table 6.3. The annotated activities were split into testing and training sets using a leave-one-out cross-validation strategy, where one match half was used for testing and the remaining halves for training. A k-Nearest Neighbour (k-NN) classification approach was used, taking the mode activity label of the closest k examples in the training set, and using the L_2 distance measure. While a more sophisticated classification approach could be used, k-NN provides quick classification and allows the different representations to be compared. Confusion matrices using $k = 10$ are presented in Figure 6.4.

From the confusion matrices in Figure 6.4, it can be seen that most activities are well recognised, however goals are often misclassified as the other activities. This can be explained by goals being less structured than the other activities, with a lot of variability possible. Defensive corners and long corners are sometimes confused with one another as the main difference is the team which maintains possession, which is not discernible from the occupancy descriptors.

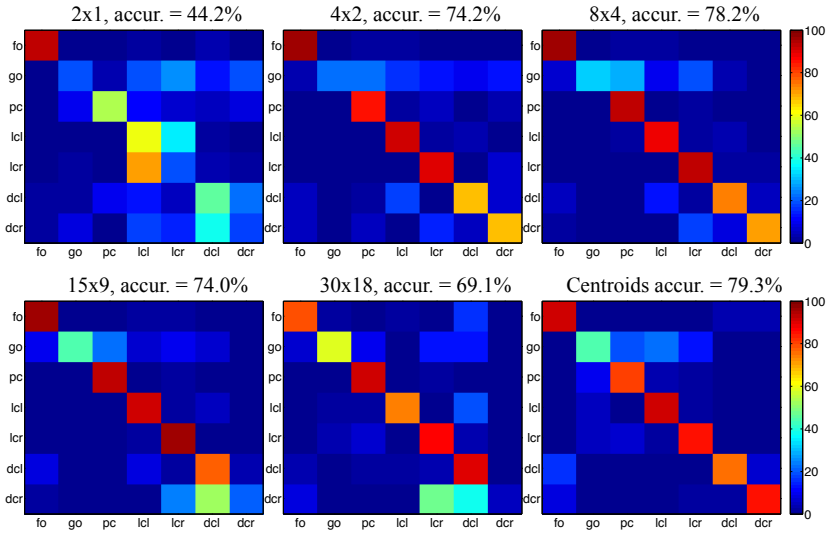


Figure 6.4: Confusion matrices for isolated activity recognition using different occupancy map descriptor sizes and the centroid representation.

The best accuracy was achieved using the centroid descriptor, with an accuracy of 79.3%, followed closely by an 8×4 occupancy map descriptor with an accuracy of 78.2%. Quantising at a coarser level was not able to distinguish between the activities as accurately, while quantising at a finer level beyond this resulted in a slightly reduced accuracy. This can be explained by players not aligning to the exact locations in the training activity templates as there is a high degree of variability in the player positions for each activity (and the L_2 distance only compares corresponding field locations between occupancy maps). A coarser descriptor has more tolerance for player position variations, and the 8×4 descriptor is able to do so while still providing discrimination between activities. The centroids outperform the team occupancy descriptors, which may be attributed to that fact that these activities can be described on a macroscopic scale (i.e. by the global distribution of the players, which is captured by the centroid, rather than the exact positions of each player as approximated with the finer descriptors). The 8×4 occupancy descriptor captures similar granularity to the centroid representation, and is also very effective in distinguishing between the labelled activities.

Despite their simplicity, it is evident that both representations can be used to recognise important game states in the presence of noise and without any tracking information. If finer behaviours of teams are to be recognised (i.e. at the level of individual players), an occupancy map representation is more appropriate, but requires a very high dimensionality feature vector (e.g. a grid of 30×18 requires 540 dimensions per frame to represent player locations to a precision of $\sim$3 m^2). In addition, when modelling longer term behaviours, occupancy map descriptors are not very compressible in the temporal domain, because they do not directly model player movements (which are smooth), but spatial occupancies that are discrete and do not vary smoothly or predictably in time.

6.5.2 Continuous Team Activity Recognition

Recognising team activities in a continuous sense is a more challenging task than isolated recognition, as events are not separated and a lot of movements and formations can appear very similar to labelled activities. This is particularly the case without knowledge of where the ball is, and in the presence of noise. In this section, the ability of the representations in retrieving team activities in a continuous domain are qualitatively demonstrated.

In Figure 6.5, centroids for a match half are displayed with ground truth labels for goals and penalty corners. It can be seen that goals correspond to regions where both teams are located close to the goals, followed by a movement to the centre of the field. A penalty corner ('PC') is characterised by team centroids being separated for a duration of time (as they move into formation and the attacking team plans their attack), followed by a convergence towards the goal when the ball is brought into play. This information can be used to quickly recognise the game state.

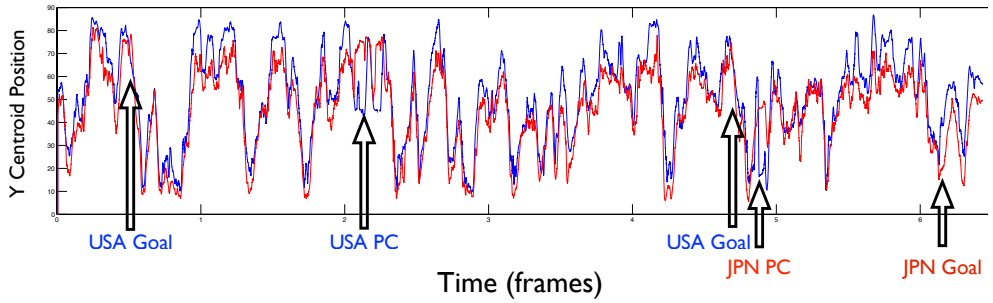


Figure 6.5: Team centroids (y-position) across match half of USA versus JPN, demonstrating that centroids provide important information for game state that can be used to assist in retrieving activities such as goals and penalty corners.

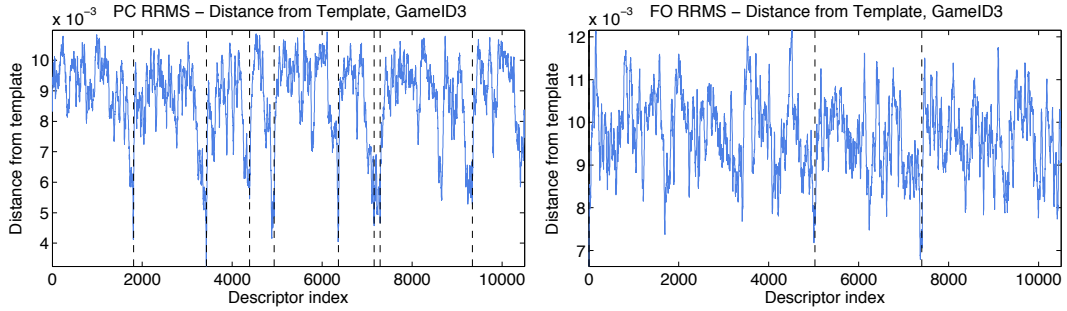


Figure 6.6: Retrieval distances for a Penalty Corner (left) and Face Off (right). The plots show the distances between the automatically extracted descriptors and the activity template to be retrieved across a match half. The ground truth activity onsets are indicated with a black-dashed vertical line. A low distance (high similarity) can be seen at the ground-truth locations.

While centroids are very useful, many team behaviours will have similar centroids, and to pick up on more specific behaviours and activities, a finer descriptor is necessary. To demonstrate retrieval, the distances between the occupancy map descriptors extracted from a game and a template of the activity of interest were calculated using a sliding window. In Figure 6.6, a 15×9 descriptor was used to recognise two different activities across a match half. A 15×9 descriptor was used as it was found that smaller descriptor sizes were often confused with non-activities when compared in a continuous domain. It can be seen that the descriptor is able to effectively locate the ground truth activity regions for a penalty corner and a face off.

6.6 Summary

In this chapter, macroscopic approaches to group behaviour alignment were presented and evaluated for the task of recognising group activities. A fully automated system was presented which is able to recognise team activities from raw player detections without player tracking or ball information. Compared to existing approaches which require continuous player tracks, the centroid and occupancy map representations proposed in this chapter are robust to missed and false detections and thus can be used with imperfect sensing system. This enables real-time group behaviour analysis to be performed without any manual pre-processing. It was demonstrated that both macroscopic representations were able to accurately recognise team activities even in the presence of noise. Recognising such coarse team activities is important for match analysis and can greatly reduce the time required by coaches to analyse a team's performance. It was also shown that occupancy maps are better suited to retrieving more specific group activities. While the macroscopic alignment approaches were applied to sports activity retrieval and classification, the proposed system can be directly applied to aligning any other multi-agent detection data captured over a fixed area, providing a very fast method of analysing and labelling spatio-temporal data.

Chapter 7

Person Re-Identification Using Formation Priors

7.1 Introduction

Recognising individuals and tracking their movements and behaviours is important in video surveillance and can allow the behaviour of a group of people to be understood. In a single camera view, this can be achieved through object tracking techniques, however, in a large space with multiple non-overlapping cameras where it is not certain which path people will take, appearance matching methods must be applied to re-identify individuals as they move between cameras. This task is termed *person re-identification*, and involves recognising an individual in different locations across a network of cameras, typically assuming that individuals wear the same clothing between sightings, as represented in Figure 7.1.

Despite the assumption that people within the environment have the same appearance from camera to camera, several complexities which arise from the envi-

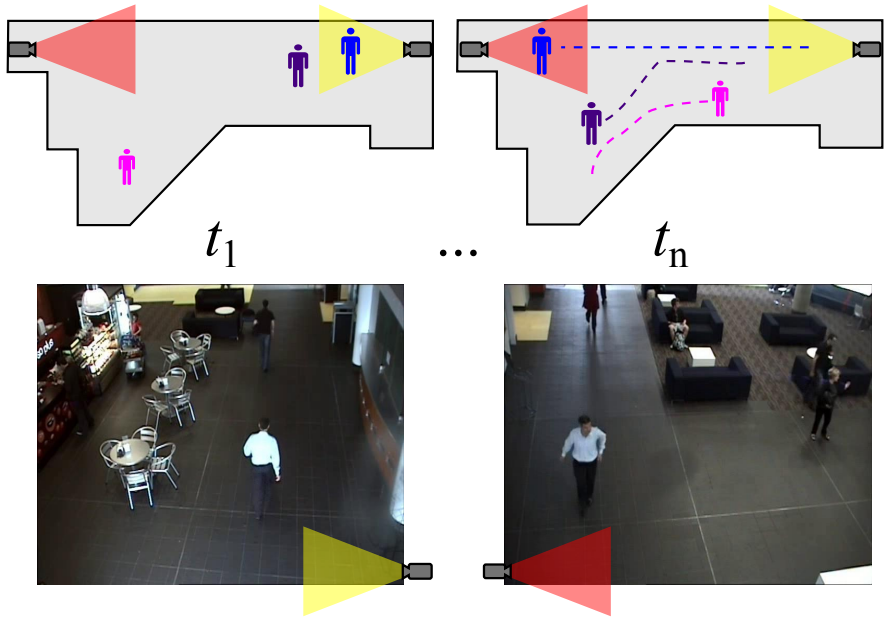


Figure 7.1: A scene at two time instants, t_1 and t_n, is represented. Person re-identification seeks to recognise the identity of a person as he/she moves between different locations, given a set of previously observed people. For example, the blue person visible from the yellow camera at t_1, later appears in the red camera at t_n. A person re-identification system should be able to reconcile this identity despite the change in appearance in the acquired video frames.

ronment make this a challenging problem, including that:

1. Subjects will often only be visible at low resolution;
2. Subjects will appear in different poses and viewpoints (e.g. front-on or side-on) as they move through the camera network;
3. The environment often contains different lighting conditions, altering the appearance of people in the space;
4. Subjects may be partially occluded.

In such conditions, traditional biometrics such as face, iris or gait generally cannot be used. Instead, models which characterise the overall appearance of a person,

or models which consist of a collection of local descriptors are used. Such models are often termed "soft biometrics" [71] and are defined as characteristics which can be used to describe, but not uniquely identify an individual. Soft biometrics include traits such as height, body build, gender, ethnicity, and characteristics which may change more frequently such as clothing colour. These features can be used to detect if a given person has been previously observed elsewhere in a network of cameras, or to search for an individual in a camera network.

In this chapter, a new database to evaluate various person re-identification features is presented, to better evaluate their performance of features in real-life surveillance conditions. Following this, it is demonstrated how group information can be used to improve performance of person re-identification in cases where appearance features alone are insufficient for distinguishing between individuals.

7.2 Related Work

The majority of existing person re-identification methods rely solely on visual information to identify individuals. In such *appearance-based* re-identification methods, approaches seek to extract a variety of global and local features from the whole-body that are distinctive and robust to viewpoint, pose and illumination changes. Colour, texture and interest point features have been extracted and classified in a number of ways.

Colour features encode appearance information of a person's clothing, hair and skin colour. They are popular for use in surveillance as they are mostly view invariant and can be sensed at a far distance from a camera. The most common method of utilising colour information is through histograms. Position information can be incorporated by splitting the image of the person into parts (e.g. in

[41, 67], histograms were extracted for the head, torso and legs) which allows matching based on colour and distribution. A "soft" binning approach [129] can be applied to compensate for illumination changes and prevent the case where similar colours are allocated to different bins. In soft histogram binning, a pixel colour value is allocated to multiple bins, weighted according to the pixel value's proximity to the centre value of each bin. Success has also been achieved using culture colours [140], which are a set of 11 colours recognised by most cultures (black, blue, brown, green, grey, orange, pink, purple, red, yellow, white), as they are less prone to illumination variation across cameras. While histograms allow for some degree of variation in colour caused by illumination by allocating a range of colours to each histogram bin, more advanced illumination compensation can be achieved using image based transformations [97], or learning a brightness transfer function [73] between cameras to compensate for the illumination variation between cameras.

Some approaches to person re-identification have used texture based features or interest points to locate and compare local patches between individuals. Gheissari et al. [50] segmented a person into regions using a triangulated graph model and compared colour and edgel information between corresponding parts of individuals, while Hamdoun et al. [61] used interest points based on a variant of SURF [14]. Both of these methods were evaluated on near frontal images, and while they were able to handle pose variations, they are unable to handle significant viewpoint variations (e.g. in a 90° rotation, the interest points would not be visible).

Other approaches extract a large number of features, and learn the most discriminative components from a training set. Gray and Tao [55] proposed an Ensemble of Localized Features (ELF), and used AdaBoost to learn the most discriminative colour and texture-based features, while Bak et al. [11] learnt the most discrimina-

tive Haar-like features and dominant colour descriptors using AdaBoost. Prosser et al. [112] reformulated the person re-identification problem as a ranking task instead of distance calculation and absolute scoring, using RankSVM to learn discriminative features. Schwartz et al. [123] proposed Partial Least Squares (PLS) reduction to project a large feature set consisting of colour, texture, and edge information into a low-dimensional discriminant latent space. Bak et al. [12] considered the multi-shot scenario where several frames of each individual are available, and proposed the Mean Riemannian Covariance Grid (MRGC) to represent people. In this method, a person is split into a grid of overlapping cells, for which covariance features [136] are extracted, and the most relevant patches to describe each individual are learnt based on variance. Unlike the boosting and ranking approaches which learn a global weighting of features across all subjects, Liu et al. [89] distributed weights to different features based on their importance in that image (e.g. colour is more informative when a person wears a textureless bright coloured shirt, while texture can be more important for a person wearing a checkered shirt).

Instead of having a training phase to learn discriminative features or regions, other methods simply extract a collection of features which are view invariant. Bazzani et al. [15] proposed a person descriptor which included a global HSV colour histogram, an 'average' texture of the person and a set of recurring textural motifs within the subject. Farenzena et al.[45] extended this work, in Symmetry-Driven Accumulation of Local Features (SDALF). They used symmetry to split a person into head, torso and legs, and added Maximally Stable Colour Region (MSCR) [47] features in the models, and achieved good view invariance. Zhao et al. [146] extracted and matched distinct salient parts based on colour and SIFT features in an unsupervised manner, and outperform SDALF, PLS and ELF. This method requires there to be unique colour or textural components within the subject set.

While these methods have demonstrated applicability in the datasets provided, it is uncertain how they would perform in different conditions, as the datasets do not allow for different evaluation conditions. Even though many of the discussed features are designed to be view and illumination tolerant, not all the datasets are able to show that this is the case, and none are able to show how the models are affected by viewing angle or illumination. Also, many approaches only look at the single image case, which is unrealistic in a surveillance network, as video is captured and available to perform foreground segmentation and allows for better selection of frames to use in the model. To overcome this, a new dataset specifically designed for person re-identification was collected and released, and is discussed in Section 7.3.

When people have very similar appearances (e.g. when wearing uniforms), the intra-person appearance variations may be greater than the inter-person variations. Therefore, additional information or context must be used to more accurately re-identify people. In all the above described methods, only appearance was used for matching as it was assumed that the camera placement and the paths that people may take between cameras was unknown. Other person re-identification methods have looked at additionally utilising the spatial layout and temporal constraints of the camera network to limit the set of candidates to be matched [72, 87, 99]. Depending on the camera network, such context is not always available, and instead other contextual information must be used. Zheng et al. [148], showed that associating groups of people instead of individuals can improve person re-identification performance, using a group descriptor which encodes visual words and their spatial relationships.

In team sports, player positions and movement are heavily linked to one another and to game context, and can be used to fill in the gaps of missed tracks caused by poor player detection. Liu et al. [90] made use of such contextual information

to improve player tracking. They extracted game context from the global and local distribution of players (to indicate which team is attacking, and situations when opposing players normally follow each other closely) to give a more accurate motion model for tracking players. Lucey et al. [95] used team centroids as a contextual feature to approximate player role in conjunction with a spatio-temporal bilinear model to clean-up noisy data. Lu et al. [93] used a conditional random field incorporating SIFT, MSER and colour histogram features to track and identify individuals in broadcast footage. Their data had sufficient resolution to detect jersey numbers (when visible) which allows for better person identification than low resolution domains, where existing methods have only looked at extracting the team of each player.

Compared to the existing approaches for person re-identification which only model the appearance of individual people, in the second half of this chapter (Section 7.4), group information is incorporated to improve person re-identification. Unlike the method of Zheng et al. [148] in which the appearance of the group was modelled, in this work, group structure is utilised and incorporated in the form of relative player positions or "roles". Also, unlike the method of Lu et al. [93] where higher resolution broadcast footage and conditional random fields that incorporated appearance and temporal information were used, this work looks at identifying players in low resolution footage (i.e. player heights of 40-100 pixels) without temporal information.

7.3 The SAIVT-SoftBio Database

To evaluate models for person recognition and re-identification, a dataset is required which consists of multiple cameras, in which the subjects appear in different poses, viewing angles and lighting conditions.

To date, researchers have used a variety of data sources to evaluate their models. Existing tracking databases have been used (e.g. [41] used a subset of PETS2006 [46]); the VIPeR (Viewpoint Invariant Pedestrian Recognition) database [55] has been used extensively (see [21, 45, 55, 66, 112]); some approaches have used the ETHZ [44] and i-LIDS [137] databases; while others have simply captured their own data (e.g. [11, 50]).

While these databases have their merits, they are limited in their ability to compare and evaluate person re-identification models in real surveillance environments. Tracking data sets typically contain few cameras and few subjects (e.g. PETS 2006 has four cameras of which only three are suitable), VIPeR only contains still images from two viewpoints for each pedestrian, ETHZ captures limited viewing angles (mostly frontal) of people, and the annotated component of i-LIDS only contains up to four images per person. While databases used in gait recognition research often contain a larger number of subjects and camera angles (e.g. the CASIA database [142] contains over 100 subjects observed from 11 cameras), they are captured in highly controlled conditions, very dissimilar to a typical surveillance environment.

Due to the limitations of existing databases that either contain only still images, few camera views, highly controlled conditions, or a lack of sufficient frames per subject, a new database is proposed together with a flexible XML-based evaluation protocol to allow for a highly configurable evaluation set-up, enabling a variety of scenarios relating to pose and lighting conditions to be evaluated. The database provides a platform from which to answer questions such as:

- What features are best for recognising the identity of a person in low resolution footage across different camera views, illumination conditions and with variable pose?

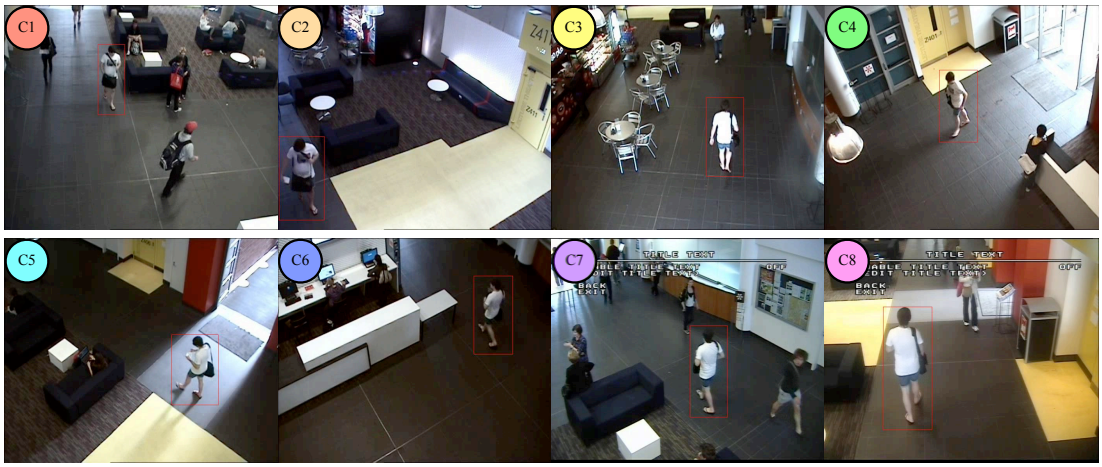


Figure 7.2: Example video frames from each of the eight cameras (C1 to C8) of the SAIVT-SoftBio database. A subject dressed in a white shirt, marked with a red bounding box, is shown in each of the cameras, highlighting the significant appearance variations (pose, viewpoint, illumination) as the subject moves through the camera network.

- How much data is necessary to build a sufficient model of a person?
- How does data from multiple views impact performance?
- Can details about pose be used to improve performance?

The utility and flexibility of the proposed database is demonstrated by using it to answer these questions with a baseline person re-detection system consisting of colour, height and texture features.

7.3.1 Database Details

The SAIVT-SoftBio multi-camera surveillance database[1] was captured from an existing surveillance network, to enable the evaluation of person recognition and re-identification models in a real-life multi-camera surveillance environment. The

[1]Available from https://wiki.qut.edu.au/display/saivt/SAIVT-SoftBio+Database

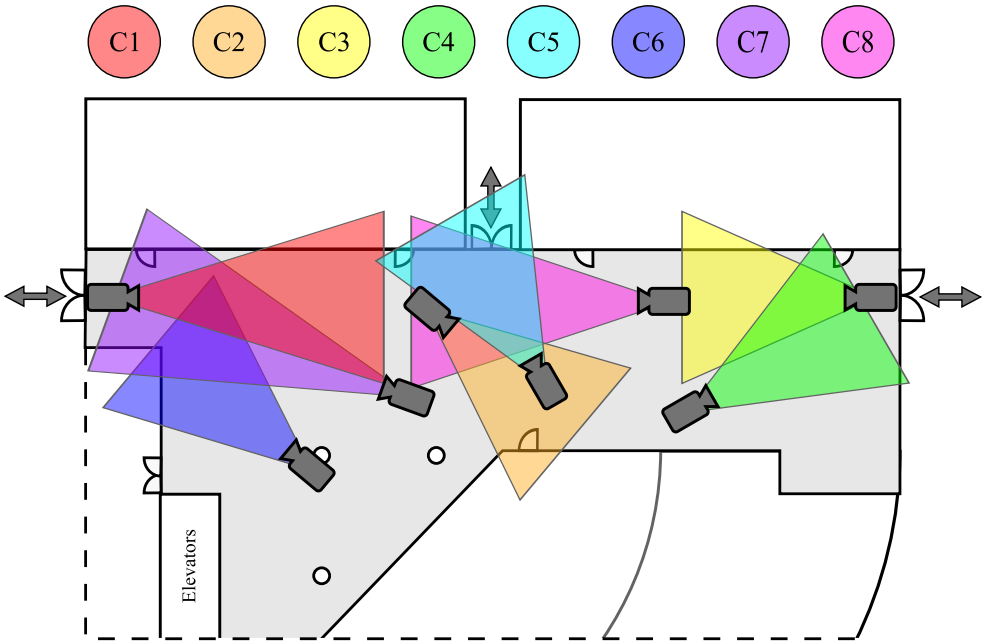


Figure 7.3: Approximate camera placement and orientation in the SAIVT-SoftBio Database. The three entrances to the building are indicated with arrows.

database consists of 150 people moving through a building environment, recorded by eight surveillance cameras. Each camera captures data at 25 frames per second, at a resolution of 704×576 pixels, and is calibrated using Tsai's method [135]. An example image from each camera is shown in Figure 7.2, with the approximate camera placement and orientation displayed in Figure 7.3. The placement of cameras is a real-life surveillance setup, and cameras have been placed to provide maximal coverage of the space (with some overlap) and observation of the entrances to the building.

The database was collected in an uncontrolled manner, so subjects can travel any route through the building. Thus, the vast majority of subjects will only pass through a subset of the camera network and that subset varies from person to person. This provides a highly unconstrained environment in which to test person re-identification models. From Figure 7.2 and 7.4, it can be seen that

Figure 7.4: Example annotations of four subjects from the Multi-Camera Surveillance Database at different locations in the camera network, where S represents the subject ID and C represents the camera number.

there is varied lighting across the different camera views, and that subjects are observed from different angles as they move through the network. To enable a consistent evaluation in such conditions, a coarse bounding box indicating the location of the subjects has been annotated (every 20th frame was annotated and intermediate frame locations were interpolated). The frames are recorded from when the subject enters the building through one of the three main doorways visible in Camera 4, Camera 7 and Camera 5/8, until they leave observation either through exiting the building or entering a lecture theatre. Any frames which are significantly occluded, have been omitted. Examples of the annotated subjects are shown in Figure 7.4

XML files are used to store information about the database to enable different evaluations to be easily performed based on which subset of the database fits the desired criteria. For each subject, an XML file is used to summarise the camera views and frame information which can be used to select subjects which fit the desired evaluation conditions (e.g. only subjects that exist in specific cameras or locations can be selected). The overall database is also summarised in an XML file, which provides information on the camera calibration data for each subject. Zones of interest can be specified to further filter the person annotations, allowing for additional conditions to be evaluated (i.e. lighting changes can either

be reduced or emphasised by only considering certain scene areas).

The database provides great flexibility in the possible evaluations that can be carried out due to the variations captured by the eight cameras. It can be used for traditional biometric identification and verification tasks, as well as person re-detection.

7.3.2 Baseline Appearance Models

To demonstrate the utility of the database, simple person models are evaluated, consisting of colour, height and texture models . The overall evaluation procedure and the steps to acquire our baseline models is displayed in Figures 7.5 and 7.6.

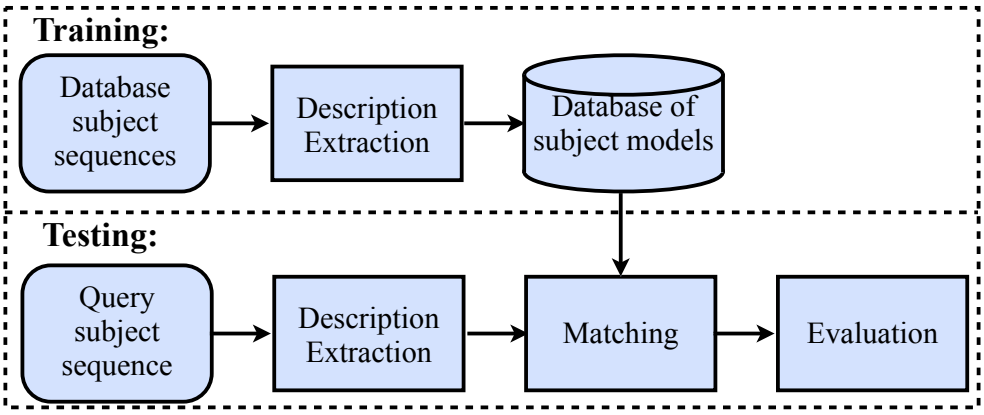


Figure 7.5: Person re-identification system evaluation flowchart

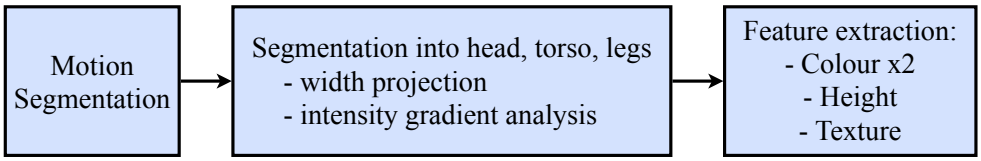


Figure 7.6: The steps involved in extracting a description of a person in the baseline system

For all models outlined within this section, a motion segmentation algorithm [42]

was used to separate the subject from the background. After extracting the foreground regions (i.e. pixels belonging to the person), the person is divided into head, torso and legs parts through horizontal projection and image gradient analysis as described in [39]. Example output from this process is shown in Figure 7.7.

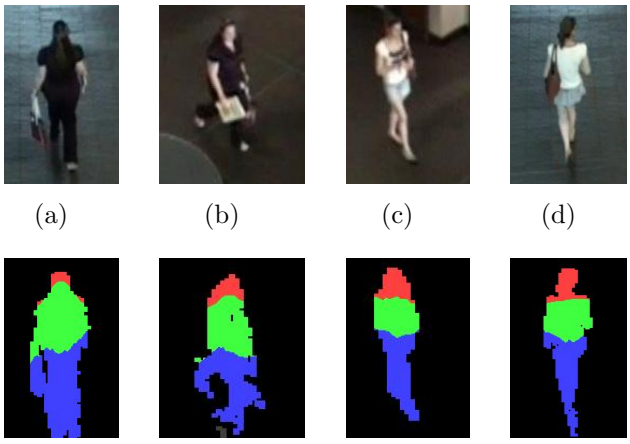


Figure 7.7: Segmenting a person into head, torso and leg regions (coloured in red, green and blue, respectively). The top row shows the input colour images, the bottom row shows the segmented silhouettes.

7.3.2.1 Colour Models

Colour information of a person is extracted by computing histograms of their head, torso and leg regions. For each of the three regions, a soft histogram of the full colour space is calculated as well as a histogram of the culture colours [140], resulting in two colour soft biometric models (soft histogram and culture colour histogram). A moving average of each histogram is calculated to incorporate multiple frames into the model.

In the soft histogram, variation in colour across different cameras is reduced through the soft-binning, where each pixel colour value is assigned to multiple

bins based on its proximity to the centre of each bin. This means that samples which lie on a bin boundary, where there is greater uncertainty, are split more evenly and prevents very similar colours from being wholly allocated to different bins.

The culture colour model quantises the image into 11 colours (black, brown, grey, red, orange, yellow, green, blue, purple, pink, white), with the aim of transforming the colours into a space less affected by illumination variations. To convert the image into its corresponding culture colour image, Gaussian mixture models (GMMs) were trained to represent each of the 11 culture colours from a set of small image patches (each containing a single culture colour). Each foreground pixel of a person is then classified into the culture colour with the greatest likelihood, and then the histograms are computed.

The histograms are normalised to sum to 1, ensuring invariance to the number of images used to build the model and the size of those images, and are compared using the Bhattacharya coefficient. When comparing colour models for two people, the similarity score is taken as the average of the three histogram region (head, torso, legs) comparisons.

7.3.2.2 Height Model

The height of a person is used as a simple descriptor height is view invariant. Other dimensions (width and depth) are dependent on the camera angle and a person's pose.

Heights are calculated using the detected positions of the head, torso and legs (which are converted into a real world coordinate scheme using camera calibration), and a soft histogram approach as described in [40] is used. Figure 7.8 shows

an example of the located head and feet points, and the points used to divide the subject into head, torso and legs.

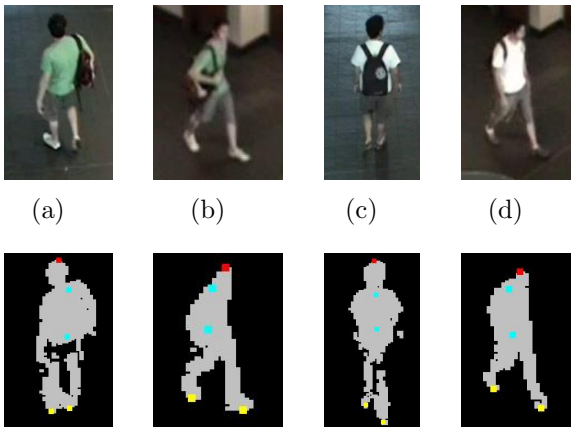


Figure 7.8: Detecting the head, neck, waist, and feet. The top row shows the colour input image and the bottom row shows the corresponding silhouette with the detected points overlaid. The head points are shown in red, feet shown in yellow, and median position of the waist and neck divisions shown in cyan.

7.3.2.3 Texture Model

To model the texture information of a person, local binary patterns (LBPs) [105] are used. The LBP is an excellent texture descriptor for its invariance to illumination, and can also be made to be rotation invariant. In this work, an LBP model consisting of 8 points with a radius of 1 pixel is used and a single texture model is extracted for the whole person. The LBP computation for an individual pixel is shown in Figure 7.9. The LBP feature vector is then computed as a normalised frequency count of each possible pixel value across all foreground pixels, resulting in a feature vector of size 256 (a neighbourhood of 8 pixels gives 2^8 possible values). Example local binary patterns which represent different textural primitives are shown in Figure 7.10.

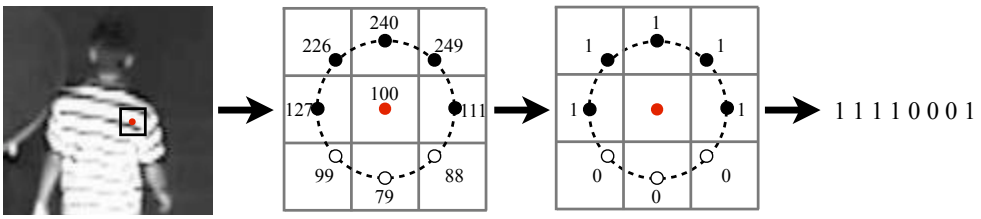


Figure 7.9: Calculating the LBP feature value for a given pixel by comparing its intensity value with those around it, resulting in an 8-bit value.

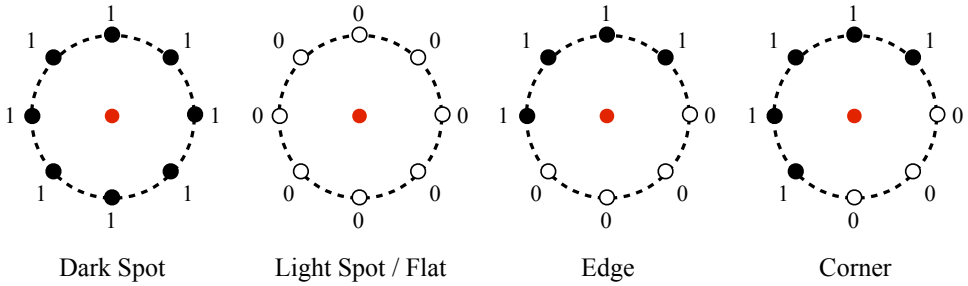


Figure 7.10: Example textural primitives represented in LBPs

7.3.2.4 Fusion

As each model (colour, height, texture) forms a weak classifier, they can be fused together to take advantage of the complementary information of each model. A weighted summation of the models is applied, so that the overall match between two people i and j is:

$$M(i,j) = \sum_{n=1}^{4} w_n \times M_n(i,j), \tag{7.1}$$

where w is the weight applied to model n (soft histogram colour model, culture colour model, height model and texture model); and $M_n(i,j)$ is the matching score for model n, between person i and j.

7.3.3 Database Usage for Feature Evaluation

To demonstrate the utility of the proposed database, the baseline soft biometrics are evaluated to see how they are affected by a variety of factors captured by this database. The results for the following evaluations are presented:

1. Effect of the number of frames considered in the models
2. Effect of viewing angle
3. Effect of the number of camera views considered in the models

Results are presented using Cumulative Matching Characteristic (CMC) curves, which represent the probability of finding the correct match in the top x matches, and Synthetic Recognition Rate (SRR) curves which represent the probability that any of the y best matches is correct, as proposed in [55]. Note that the number of subjects present in each evaluation is not consistent as only subjects that match the criteria set out for the given evaluation are used. As the database is unconstrained, different numbers of people appear in different cameras, leading to this variation.

7.3.3.1 Effect of Number of Frames Used in the Model

As a person moves through the environment, their sensed appearance will change according to the camera and ambient conditions. By considering more frames it is expected that more of this variation will be incorporated in the models. Results for this evaluation are presented using SRR instead of CMC curves, as they better represent the difference with the variable number of subjects (as the number of frames for modelling are increased, less subjects are available which fit this criteria

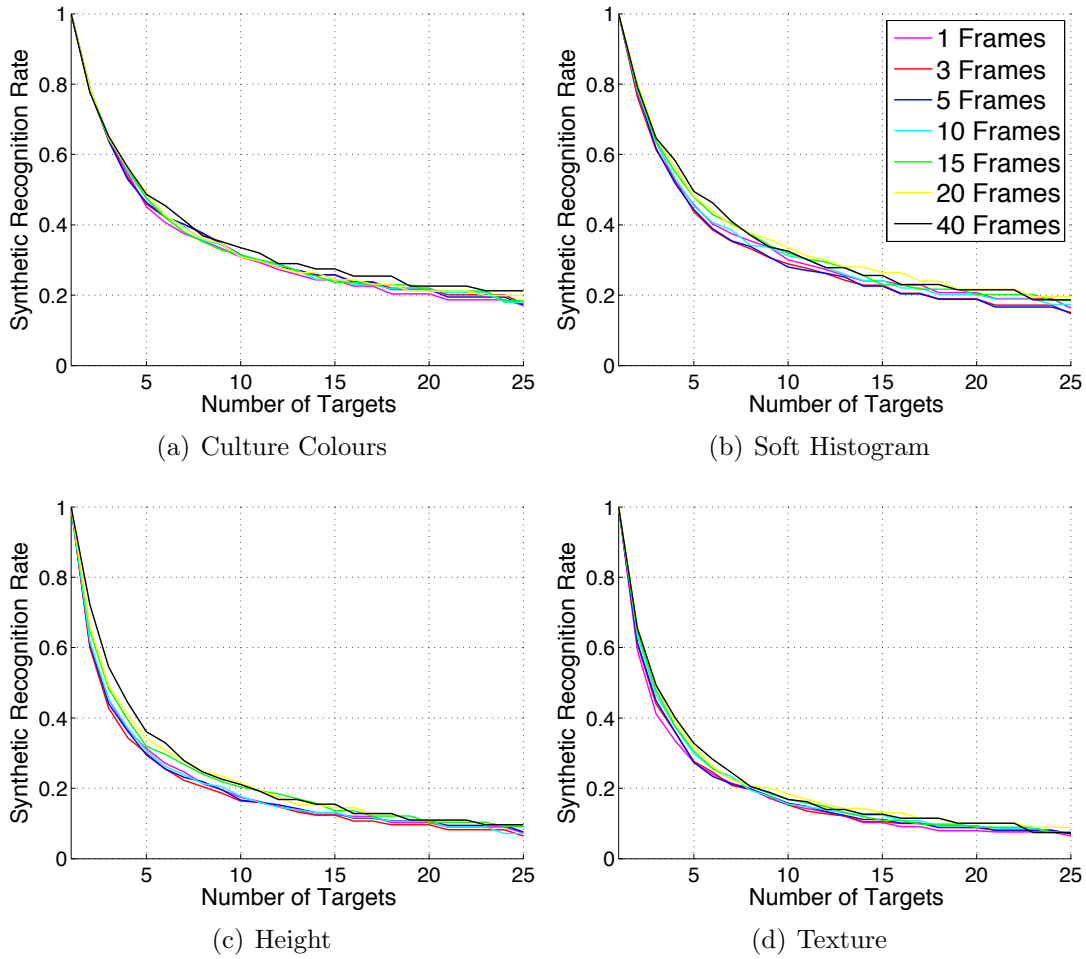
in the database). In Figure 7.11 and Table 7.1, a slight improvement is observed when considering more frames in the models (SRR values generally increase as more frames are considered, with the best performance always obtained using 20 or 40 frames). Sometimes a slight decrease is observed which may be caused by noise being incorporated in the models, for example due to segmentation errors or strong lighting variations. While generally only a small improvement is gained, having a dataset with many frames allows for motion segmentation to be performed, so only pixels belonging to a person will be incorporated in the models. Having multiple frames available for modelling a person is more representative of a realistic scenario (surveillance is captured as video), and with more frames available, certain criteria can be applied to filter out frames detected to be of poor quality due to poor segmentation or illumination as shown in Figure 7.14.

	5 targets				**10 targets**			
#Fr	CC	SH	H	T	CC	SH	H	T
1	45.1	45.7	31.3	27.4	30.8	30.1	17.4	15.3
3	46.0	43.7	29.9	27.7	30.8	28.9	16.4	15.3
5	46.3	44.3	29.6	27.40	31.4	28.0	16.7	15.8
10	47.6	45.4	30.7	29.8	31.3	31.2	17.7	15.5
15	47.5	47.9	31.9	30.6	31.5	32.0	20.3	16.9
20	49.3	48.1	34.0	32.0	30.9	33.6	21.5	18.4
40	48.7	49.5	36.0	32.8	33.5	32.5	21.0	16.8

Table 7.1: Synthesised recognition rates (%) from Fig 7.11 for 5 and 10 targets with increasing number of frames. The best #frames is shaded for each model. [Models: CC = culture colour, SH = soft histogram, H = height, T = texture]

7.3.3.2 Effect of Viewing Angle

To evaluate the effect of viewing angle, the data used in testing and training was limited to two camera views which captured similar viewing angles of a subject (Camera 3 and 8), and two camera views which capture dissimilar views of a subject (Camera 5 and 8). Assuming that subjects walk straight through the



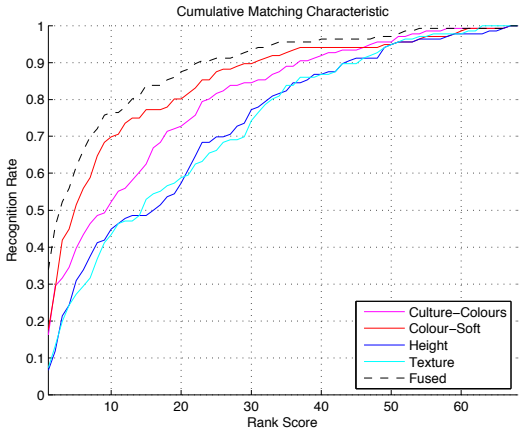
(a) Culture Colours (b) Soft Histogram

(c) Height (d) Texture

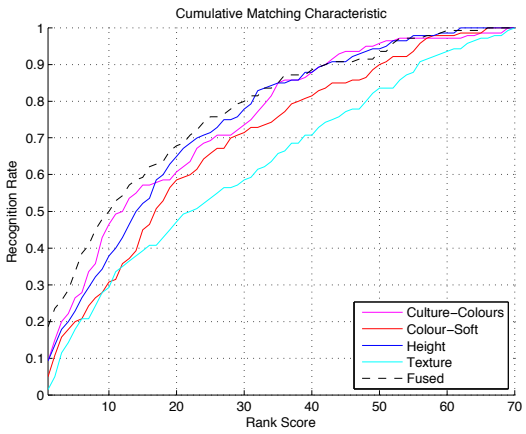
Figure 7.11: Effect of number of frames used in the model when building models from a single camera view. All camera views are considered in this evaluation, with gallery and probe models trained off separate views. (See Table 7.1 for values at 5 and 10 targets)

building and do not turn around (i.e. subjects walk left to right or right to left in the building diagram), it can be seen in Figure 7.3 that similar viewing angles will be obtained in Camera 3 and 8 and dissimilar angles will be obtained in Camera 5 and 8 as demonstrated in Figure 7.12 (c). This assumption holds true for all subjects in the database, except one which was excluded from this evaluation.

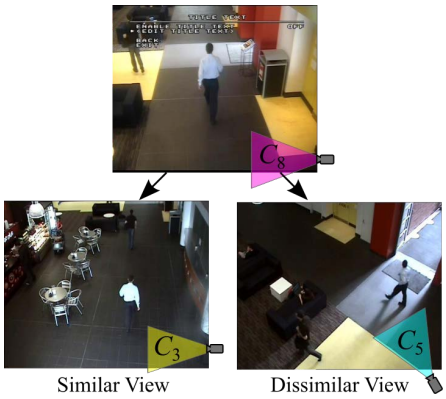
From the results presented in Figure 7.12, it can be seen that all models degrade in performance with dissimilar views (recognition rates in Figure 7.12 (b) are lower



(a) Similar view (Cam 3 and 8)



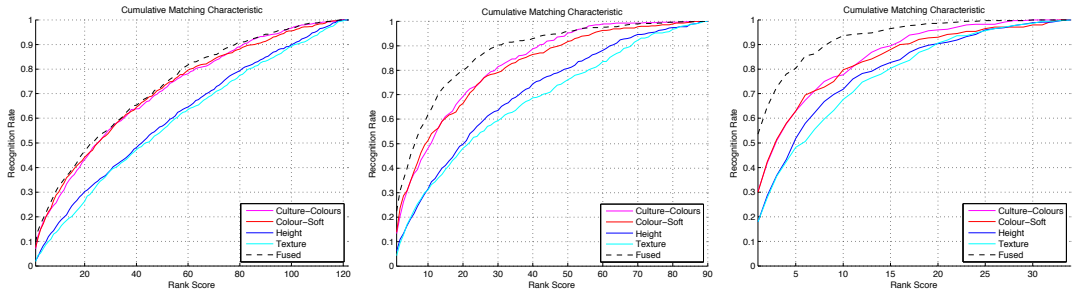
(b) Dissimilar view (Cam 5 and 8)



(c)

Figure 7.12: The effect of viewing angle mismatches in training and testing. Evaluations consider gallery and probe models trained on separate views, with models built off 20 images. (a) shows CMC plots where testing and training models contain similar viewing angles, while in (b) testing and training models are built from dissimilar viewing angles. (c) displays example frames of a person in the selected similar and dissimilar camera views.

than in Figure 7.12 (a)), except for height which works similarly in differing viewing conditions. For example, Height Rank-10 performance only degrades slightly from 45% to 38%, while Colour-Soft degrades significantly from 70% to 31%, suggesting that height is more tolerant to view variation. This is expected, as height does not change from different viewing angles while colour and texture of a person may be different from the front/side/back. The full soft colour model outperforms culture colours in similar viewing angles as seen in Figure 7.12 (a),



(a) 1 View (b) 2 Views (c) 3 Views

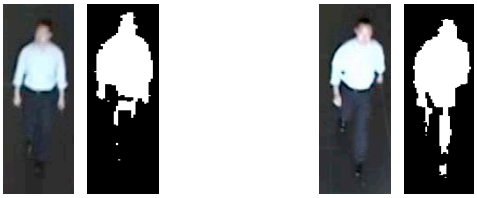
Figure 7.13: CMC plots for colour, size, texture models, trained and tested on 1, 2 and 3 camera views using 20 images each.

but culture colours perform better than full colour in exclusively different viewing angles as in Figure 7.12 (b). This suggests that culture colours or other heavily quantised colour spaces are more stable than full colour histograms in varied viewing conditions. The degradation in performance in the colour and texture models may be attributed to the fact that many of the subjects appear different from the front, side and back due to items they are carrying (backpacks, shoulder bags) and their clothing (such as open jackets). Considering all viewing angles, as in Figure 7.13, it can be seen that colour features are the most discriminative, and the most robust across all variations.

7.3.3.3 Effect of the Number of Viewpoints

In Figure 7.13, plots are presented for models trained on 1, 2 and 3 views. All cameras were considered, with mutually exclusive views used in gallery and probe models. All models were built using 20 frames. Colour models consistently outperform the height and texture models, and all models improve as more views are used to train the models. This is expected, as including different viewing angles in the model incorporates more information in the models and thus better represents the person's overall appearance.

The superior performance of the colour models compared to height is expected, as there is more variation in colour, and heights will only differ by a few centimetres between subjects. Also, the height model is more affected by errors in segmentation (both of foreground pixels and segmentation into head, torso and legs). Small errors in the silhouette can result in a difference of a few centimetres or more, depending on where in the image the subject appears. While the colour biometric is also susceptible to segmentation errors, the colour models are less affected, except where segmentation errors result in large portions of the person not being visible (e.g. their legs or torso are not detected, as in Figure 7.14 (a)), or a large portion of the background being included in the model. The poor performance of the texture models may be caused by poor resolution which results in blurring of texture, and the lack of textural information in the majority of subjects. However, texture performs fairly consistently in differing conditions. In all cases, a fused model outperforms all individual models, as the complementary information from each model combined gives greater discrimination between people.



(a) Poor segmentation (b) Better segmentation

Figure 7.14: An example of (a) poor segmentation and (b) better segmentation. Poor segmentation can result in missing body parts and reduce performance of the models. When using video footage with many frames available per subject, frame selection criteria can be used to filter out the poorly segmented frames.

7.4 Using Group Information

In a surveillance setting, the difficulty of re-identifying a person exponentially increases proportional to the time the person was last seen. This is because the number of possible options a person has increases over time, which in terms of permutations, quickly increases to infinity. For example, a person seen on a street corner at time t_0 has numerous options in terms of what they are doing next. They could potentially walk down the adjacent streets, hop on a train/bus/taxi, change clothing, or even just wait on the corner. After ten seconds, a good estimate of where that person is could be obtained. This is due to the limited options a person could take in that time, in addition to the scene remaining somewhat familiar. One minute later, the number of options increases and makes this task much more difficult. Now, ten minutes later this task is near impossible (unless the person has not moved) as the number of possible permutations explodes. However, if that person was part of a large group of people, the likelihood of that group being confused as another group is much lower. The search space of re-identifying a person within a group can thus be greatly reduced by just choosing the most likely candidate within that group of people - instead of the huge number of possibilities if the person was alone. As humans are social beings, most realistic settings have people moving as a group rather than an individual. In this section, the group dynamic is leveraged to improve person re-identification.

Although the above assumption is obvious, the big bottleneck which has restricted research in this area is the collection and annotation of large amounts of group data. As such, nearly all the work in person re-identification has solely focussed on modelling a single person (apart from the work of Zheng et al. [148]). With the influx of vision-systems being applied in the sporting domain due to the commercial applications, sporting datasets provide a perfect test-bed for investigating different approaches for person re-identification. They provide similar scenarios

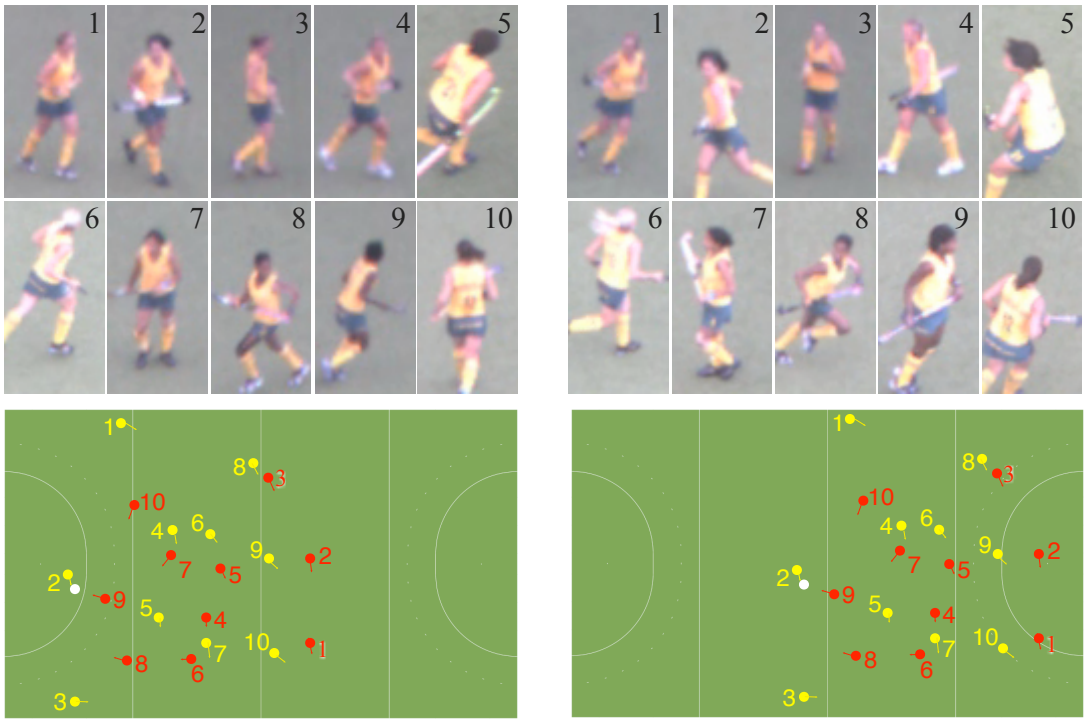


Figure 7.15: The players of a sports team are represented at two time instants, (a) and (b). It can be seen that the player appearances vary significantly between the two time instants, in terms of illumination, viewing angle and pose, and it is difficult to distinguish between the players because they are wearing the same uniform. While appearance alone is ambiguous, the structure of the team, represented in the bottom half of the figure, remains similar. This group context can be fused with appearance features to improve person re-identification.

to surveillance environments, where resolution is generally low and people may have similar appearances (e.g. a group of school children in uniform).

In this work, the task of recognising the identity of people with very similar appearances is undertaken using team sports video data (see Figure 7.15). This provides a constrained environment to evaluate person re-identification models as there are a fixed number of subjects and a well defined area where they can be observed. It is also extremely challenging, as the video is captured in an outdoor environment by multiple cameras, which results in significant illumination variations between observations of each subject, the image resolution of players is low, which makes digit recognition on the jerseys near impossible, and player poses vary significantly. Such a dataset allows person re-identification models to be evaluated in real-life conditions, and allows the use of group context to be evaluated. Using this dataset, existing state-of-the-art appearance-based features are evaluated. Then, group context in the form of player roles is used to help disambiguate between people with similar appearances, and this contextual information is shown to improve person re-identification performance.

7.4.1 Evaluation Overview

7.4.1.1 Dataset

To evaluate person re-identification using group context, team sports video data is used as it provides a real-life outdoor environment to evaluate person re-identification models with group context, and a fixed number of subjects are visible over a long duration with repetitive behaviours and structure. Typical challenges of person re-identification are present such as variations in subjects' orientation (e.g. viewed from front/side/back), pose (standing straight, crouch-

Figure 7.16: Example image patches of a single player, captured at different times and locations on the field are shown. A wide degree of appearance variation in terms of illumination, viewpoint, and pose is apparent.

ing), resolution, and illumination. Examples of these variations for a single player are shown in Figure 7.16.

The field hockey test-bed described in Chapter 5 was used to provide a dataset of player images as well as group context. While the camera test-bed provides complete coverage of the field, the task of person re-identification was considered instead of tracking, to observe how group context can assist in person re-identification. Image patches were automatically extracted using a state-of-the-art real-time person detector [28], and the images were scaled to a fixed size of 96×50 pixels. Ground truth player positions and their identity were manually labelled, and from this the image patches were assigned a player identity based on the closest labelled player position.

After detecting the player positions and extracting their image patches, the player locations from all eight cameras were merged based on proximity, and the team or "group" was assigned based on a colour histogram of the player in the LAB colour space. This essentially provides a top down view of the match with player positions and their team.

To enable the learning and evaluation of group context, the roles were trained from a set of matches, totalling 25,000 frames. The evaluation of roles and person re-identification was performed on a held out match, where the team dressed in

yellow tops and green skirts was considered. Example images for 10 of these players are shown in Figure 7.15. Two parts of this match were annotated to evaluate role assignment and person re-identification (consisting of 3893 frames and 8838 frames respectively). The person re-identification evaluation was performed on a set of 94 images automatically extracted for the 13 players of the team.

7.4.1.2 Appearance Features

In a domain where people are only visible at low resolution, and in different orientations and poses, features which can be extracted from a distance and which are invariant to these variations are desirable for person re-identification. In a sports domain, players within a team have very similar appearances as they wear the same uniform. Given high resolution imagery of each player, they could be uniquely identified based on their facial characteristics or jersey numbers, however these features are not visible due to insufficient resolution and motion blur. Texture, which has been used in a number of existing approaches, does not provide any information for distinguishing between individuals in this context as they all wear the same uniform. Height and body shape varies slightly between players, but due to the unconstrained pose and orientation, it is not possible to accurately extract these. Traits which are visible at long range and which may distinguish between players include hair, skin and shoe colour, and therefore descriptors which encode colour and spatial information were used. Two types of appearance features were evaluated for describing each person: Symmetry-Driven Accumulation of Local Features (SDALF) and region covariance descriptors.

In the SDALF approach proposed by Farenzena et al. [45], a person is split into their upper body and legs based on symmetry, and features are extracted and matched between these parts. In this work, illumination and colour variations between the cameras were normalised by scaling the mean illumination of each

image patch to a fixed value and applying histogram equalisation. After this, a weighted histogram and maximally stable colour region (MSCR) features were extracted for each body part (the recurrent highly structured pattern features normally in SDALF were excluded, as players are dressed in the same uniform and will not vary texturally). By splitting a person into body parts, coarse correspondence can be gained when matching features.

In the region covariance descriptor [136], information is encoded about the variance and correlations of features within an image region. To represent the appearance of players, a collection of such descriptors was calculated for each player image by splitting the image into a grid of regions, and describing each region by a covariance matrix, $\mathbf{C}_R$. The covariance matrix of a region R consisting of N pixels can be computed as:

$$\mathbf{C}_R = \frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{z}_k - \boldsymbol{\mu})(\mathbf{z}_k - \boldsymbol{\mu})^T, \tag{7.2}$$

where $\mathbf{z}_k$ is the d-dimensional feature vector used to represent a pixel ($k = 1...N$), and $\boldsymbol{\mu}$ is the mean of the pixel feature vectors in the region.

The features used to represent each pixel were the x and y spatial coordinates of the pixel and the intensity values in each colour channel (R,G,B):

$$\mathbf{z}_k = [x, y, R_{xy}, G_{xy}, B_{xy}]. \tag{7.3}$$

Gradient features were not used as they were found to decrease performance, likely due to large variations in player pose.

Corresponding regions between two images were then compared using the following distance measure [48]:

$$\rho(\mathbf{C}_1, \mathbf{C}_2) = \sqrt{\sum_{i=1}^{d} \ln^2 \lambda_i(\mathbf{C}_1, \mathbf{C}_2)}, \tag{7.4}$$

where $\{\lambda_i(\mathbf{C}_1, \mathbf{C}_2)\}_{i=1...d}$ are the generalised eigenvalues of $\mathbf{C}_1$ and $\mathbf{C}_2$, computed from:

$$\lambda_i \mathbf{C}_1 x_i - \mathbf{C}_2 x_i = 0, \text{for } i = 1...d, \tag{7.5}$$

where $x_i \neq 0$ are the generalised eigenvectors.

To get the overall distance between two images, a weighted summation of the cell region distances is computed:

$$\text{Distance}(Subject_A, Subject_B) = \sum_{m=1}^{M} \frac{\rho(C_{A,m}, C_{B,m})}{\sigma_{A,m} + \sigma_{B,m}}, \tag{7.6}$$

where m corresponds to the cell region number up to the total number of cells, M, and $\sigma_{A,m}$ and $\sigma_{B,m}$ are weighting parameters for cell region m for subject A, and B respectively. The σ terms are called "variance" [12] and are calculated as:

$$\sigma_{a,m} = \frac{1}{S-1} \sum_{s=1;s\neq a}^{S} \rho^2(C_{a,m}, C_{s,m}), \tag{7.7}$$

where a is the subject index, and s corresponds to all the other subjects in the database, and S is the total number of subjects. In this way, each cell is weighted according to its "variance" or how much it varies across subjects. This gives greater weight to more discriminative cells (i.e. the regions which differ most from other subjects and have a greater σ value, represent regions that are more discriminant).

7.4.2 Role Assignment

By segmenting groups of people in an environment and identifying the type of group, roles within the group can then be identified to improve identification. A bottom-up diagram of the proposed approach is shown in Figure 7.17. For example, to locate a missing child who was last with their parents and two other siblings (i.e. a family of five) one strategy could be to go through the footage

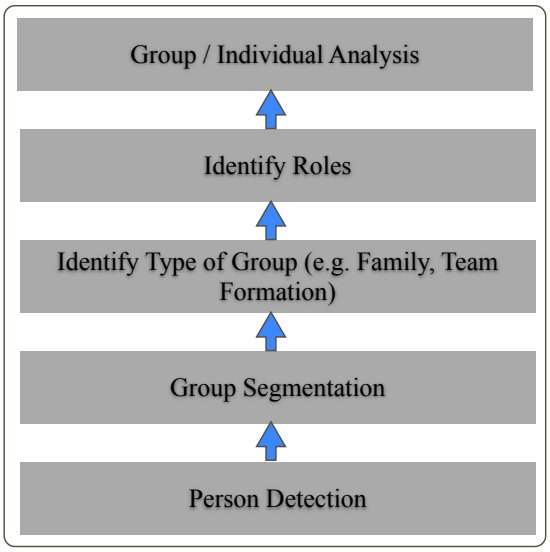


Figure 7.17: Group information can be used in a bottom-up approach to improve individual and group behaviour analysis within groups

to determine the last moment they were observed together and track from that moment to help locate the child. Rather than looking for the child alone, who could have a similar appearance to thousands of other children in the environment, knowing the appearance of his whole family and searching for groups with 2 adults and 3 children with a given appearance could limit the options and assist in finding the child much more efficiently. The type of group and roles depend on the context. Because surveillance data is limited, the analysis performed in this chapter is restricted to team sports where roles are defined based on structure and relative positions within the structure.

In the majority of team sports, the coach or captain designates an overall structure or system of play for a team. In field hockey, the structure is described as a formation involving roles or individual responsibilities. For instance, the 5:3:2 formation defines a set of roles R = {left back (LB), right back (RB), left halfback (LH), centre halfback (CH), right halfback (RH), inside left (IL), inside right (IR), left wing (LW), centre forward (CF), right wing (RW)}. This is shown in Figure 7.18. While players may swap roles throughout a match, they will

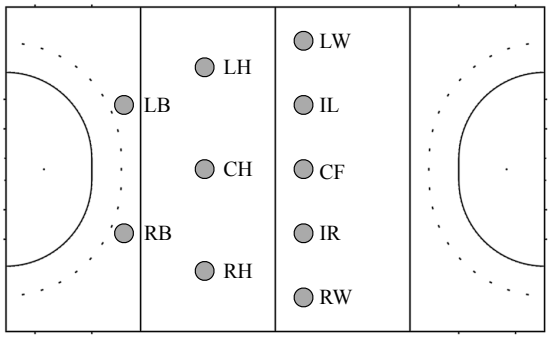


Figure 7.18: In field-hockey, players move as a formation, with each player in the team being assigned a role or responsibility. Given that the locations of all the individuals can be sensed, the role that each player takes within the formation at any instant in time can be estimated and used to assist in identification.

predominantly play in one role, and hence roles can be used as a contextual feature for identifying players. While spatial relationships are used to distinguish roles in team sports, for other types of groups roles can be inferred based off other features (e.g. height, gender, age can be used in family member classification [36]).

For training and evaluation purposes, the roles of the players were manually labelled. To give an indication of how well roles correspond to player identities, the frequency that each player identity was assigned to the manually labelled roles is presented in Figure 7.19 as confusion matrices. This was sorted so that the players most likely to play in each role appear on the diagonal. From these matrices, it is apparent that roles provide information towards player identities.

To assign players to roles automatically, a similar procedure to that proposed in [95] was adopted. First, the location of all the players on the team was detected, and then each player's position was mapped to a role within the formation. Given an initial ordering of the 10 field players of a team, $\mathbf{p}_t = [x_1, y_1, x_2, y_2, ..., x_{10}, y_{10}]^T$, at time instant t, the goal is to find the 10×10 permutation matrix, $\mathbf{x}_t$, which re-arranges the players into role order: $\mathbf{r}_t = \mathbf{x}_t \mathbf{p}_t$. The permutation matrix is a binary matrix, and if element $\mathbf{x}_t(i, j) = 1$, it indicates that player i is assigned to

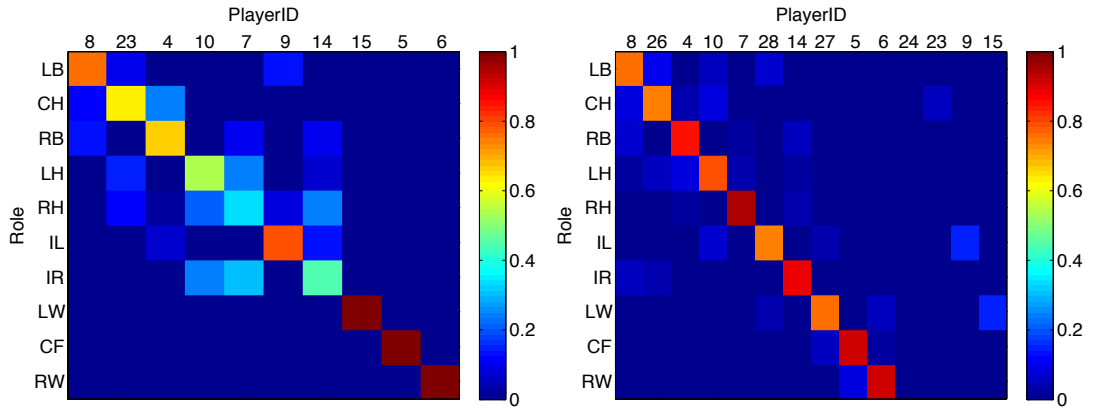


Figure 7.19: Distribution of roles to player identities from the manually labelled player roles and identities for part 1 and part 2 of the match. It is apparent that players tend to play one main role, and sometimes swap to similar (neighbouring) roles. It can also be seen that in the second half, players have been substituted so their role is replaced by another player (e.g. $9 \rightarrow 28$, $15 \rightarrow 27$).

role j. By definition, every row and column in this matrix must sum to one (i.e. each player is assigned to one role).

The role assignment task is formed as an optimisation problem where the goal is to minimise the L_2 reconstruction error:

$$\mathbf{x}_t{}^* = \arg\min_{\mathbf{x}_t} \|\hat{\mathbf{r}} - \mathbf{x_t}\mathbf{p_t}\|_2^2. \tag{7.8}$$

This is a linear assignment problem where an entry $C(i,j)$ in the cost matrix is the Euclidean distance between role locations:

$$C(i,j) = \|\hat{\mathbf{r}}(\mathbf{i}) - \mathbf{p_t}(\mathbf{j})\|_2 \tag{7.9}$$

To solve the assignment problem, the most similar prototype formation, $\hat{\mathbf{r}}$ is found from a set of 25,000 labelled frames of field hockey data, based on the mean and covariance of the team's formation in the current frame. Then the optimal permutation matrix is found using the Hungarian algorithm [79].

To evaluate the automatic role assignment performance, the results were com-

pared against manually annotated role labels and this is presented in Figure 7.20 as a confusion matrix. A major diagonal is evident indicating good performance, but sometimes the roles are wrongly classified due to ambiguity in the formation or roles, particularly in the midfield (RH, IL, IR). The overall accuracy of the automatic role assignment method was found to be 66.0%.

Since it is not known in advance which roles a player will take throughout a match, it is assumed that each player has an 'assigned role' as shown in Table 7.2. Given an assigned role, the most likely player ID can be estimated. Due to player substitutions and multiple players being able to take each role, it is evident that using role context alone is insufficient for identifying players. In addition, it is unknown which players are on the field at any time, and so there will be ambiguity in which player is playing. It is expected that appearance features will improve results.

7.4.3 Experiments

To evaluate person re-identification performance and the proposed group context using player roles, the performance of each feature was evaluated: 1) appearance features alone, 2) role alone, and 3) combined role and appearance information.

LB	CH	RB	LH	RH	IL	IR	LW	CF	RW
8	23	4	10	7	9	14	15	5	6
	26				28		27		

Table 7.2: Player IDs assigned to each role

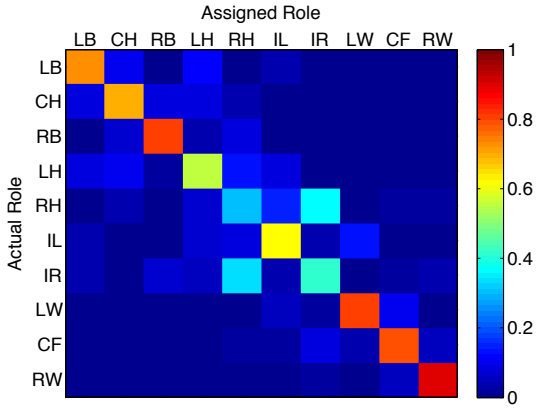
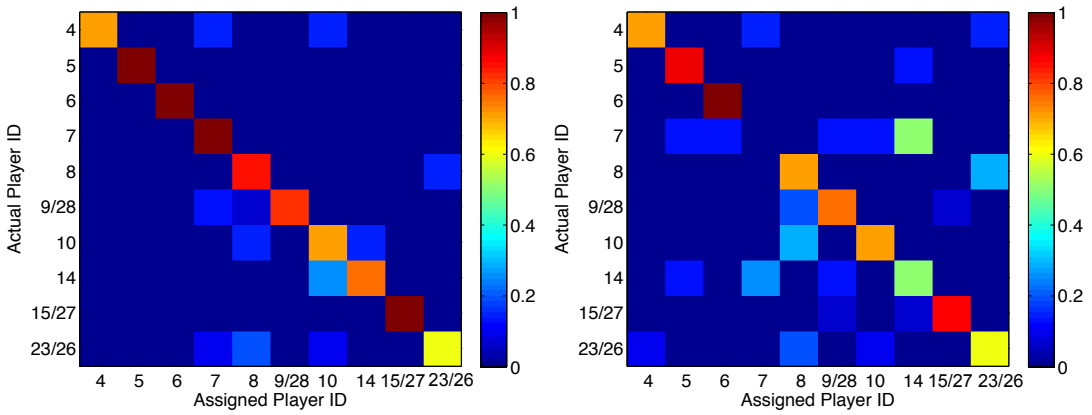


Figure 7.20: Accuracy of automatic assignment of roles (66.0%)

7.4.3.1 Identification using Roles

Given prior knowledge of which roles corresponds to which players (as in Table 7.2), identification can be performed directly on any member of a group formation without a gallery or training set. In Figure 7.21, identification results for all 94 images of the evaluation dataset, using role only are displayed. Results are shown for both perfect role extraction (i.e. roles manually annotated by an expert), and automatic role extraction. Note that due to player substitutions, it is not possible to distinguish between the substituted players (23/26, 9/28, 15/27) using role alone.

When the role to player correspondences are not known before-hand, person re-identification can be performed by comparing roles of the testing subjects to those in the gallery set. This requires a distance measure of how similar or different a role is to another role. To get a distance measure between any two roles, the average confusion matrix from the automatic assignment accuracy was used (see Figure 7.20) because roles which are easily confused must be similar. These values were converted from measuring similarity to a distance measure by subtracting the average assignment probabilities from a matrix of ones (and hence similar roles will be given a lower distance comparison measure).



(a) (b)

Figure 7.21: Accuracy of person identification using (a) manually labelled roles = 84.5% and (b) automatically assigned roles = 67.4%

.

7.4.3.2 Comparing Features for Identification

Cumulative Matching Characteristic (CMC) curves [55] were used to present results. Each point is calculated as the cumulative probability that the actual subject of a test measurement is among its k top matches (where k is called the rank). For example, the Rank-1 value indicates what proportion of the time a player is the closest match to itself, while Rank-5 indicates how often the correct player is within the top 5 matches to itself.

The evaluation was performed on a set of 94 images automatically extracted for the 13 players of the team. Two randomly selected images of every player were selected from the database (one for the gallery, and the other as part of the test set). The gallery and test set were then matched, and the results of 100 experiments were averaged to produce the results. The SDALF features, weighted covariance grid ("weightedCG"), and role assignments (automatically generated from the player position within the formation) were compared, as well as appearance features combined with roles using a weighted summation. These results are presented in Figure 7.22.

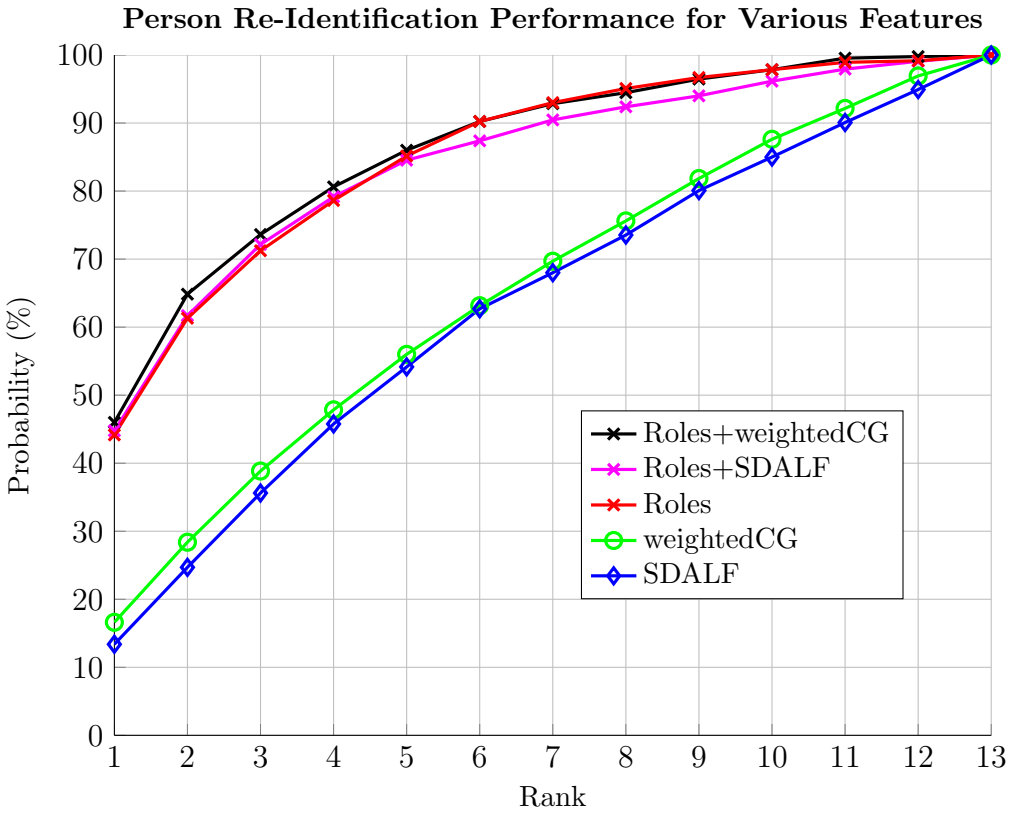


Figure 7.22: Cumulative Matching Characteristic curves for each of the person re-identification features, displaying the probability that the correct subject is within the top-k matches to itself, where k is the rank.

In Figure 7.22, it can be seen that both appearance features, SDALF and weighted covariance grid features, perform similarly poorly. This is expected as the players are all dressed very similarly, and appearance features in low resolution footage are insufficient to distinguish between the players. The weighted covariance grid slightly outperforms SDALF, and this may be due to the discriminative weighting of the cells, potentially picking up on regions that differ such as hair and shoe colour. In comparison, roles are able to distinguish players very well, and a minor improvement was gained by additionally incorporating the appearance features.

7.5 Summary

In this chapter, two major contributions to the field of person re-identification were presented - 1) a new database for the evaluation of person re-identification models in real surveillance conditions and 2) the use of group information as a contextual cue to improve person re-identification.

With the presenting of the database, a set of baseline models were used to show how this new database can be used to better evaluate person recognition models in variable real-world conditions. In particular, it was demonstrated how this dataset can be used to evaluate a number of scenarios related to number of frames, number of cameras and viewing angles which can only be evaluated with a database consisting of a large number of subjects in a variable and unconstrained environment. With the baseline models, it was shown that colour models perform better across all viewing angles as there is greater discrimination in the models compared to height and texture. When considering different viewing angles exclusively, height was found to be quite stable, with colour and texture seen to be more view specific, as many subjects in the dataset appear different from the front, side and back due to carrying of objects (e.g. backpacks) and clothing characteristics (e.g. open jacket). It was also observed that culture colours (a quantised set of 11 colours) were slightly more stable than full colour histograms, suggesting that a heavily quantised learned colour space is preferable when encountering view mismatch.

Person re-identification is very difficult in low-resolution video footage, especially when people wear similar clothing which limits the usefulness of traditional appearance-based approaches. To circumvent these issues, the use of *group information* as a contextual feature was proposed to aid in the re-identification of a person. To encode group context, a linear mapping function to assign each per-

son to a "role" or position within the group structure was proposed. Then, the appearance and group context cues were combined using a weighted summation. This was demonstrated to improve performance of person re-identification in a sports environment compared to appearance based-features, and could further be extended to more specific roles that appear in surveillance environments (e.g. family roles)

Chapter 8

Conclusions and Future Work

8.1 Summary of Contributions

The past few years have seen a deluge of visual and spatio-temporal data become available for analysing group behaviours. While a lot of work has been involved in sensing and tracking agents, a major bottleneck that has restricted large-scale group behaviour analysis has been the complexity in dealing with multi-agent trajectory data. Various sources of misalignment occur in group behaviour data that make large-scale group analysis difficult, including frequent role swaps between individuals of a group, substitutions or changed identities within the group, comparisons of different groups and missing or noisy data. This has prevented large-scale analysis of group behaviours, and in particular fine-grained analysis and analysis of noisy data.

In this thesis, role information was utilised to align multi-agent data, and enables the characterisation of groups and large-scale analysis of their behaviours. No other research has worked with this amount of multi-agent data before, and a

major contribution in this thesis was the development of alignment methods to enable large-scale analysis of group behaviour data. Macroscopic and microscopic approaches were proposed for aligning group behaviour data and their utility was demonstrated for analysing team behaviours on millions of frames worth of data in professional soccer and field-hockey analysis. Group context was also demonstrated in improving person re-identification, which can be used to locate individuals within group situations, correct tracking results, and facilitate group behaviour analysis.

The specific contributions of this thesis can be summarised as:

(i) A major contribution was the development of an alignment procedure based on roles which enables large-scale analysis of group behaviour data. In the proposed role representation, the vector representing the location of each agent of a group at any time instant is re-ordered to a template, to provide a consistent and compact representation across large datasets. This overcomes frequent role swaps which cause high variance in the data, and provides a more compressible signal for performing clustering and analysis of multi-agent data.

(ii) Various representations were proposed for representing a team's formation, including a completely automated procedure based on minimum entropy data partitioning. In this method, the underlying formation of a group can be discovered by minimizing the entropy of a set of player roles, disentangling the overlapping player distributions into distinct role distributions.

(iii) A host of new methods to characterise and compare group behaviours from large spatio-temporal datasets, made possible through alignment were proposed including:

  – Discovery, visualisation and clustering of team formations

- Player analysis using group context
- Characterisation of team style from spatio-temporal data and predicting of future playing styles
- Analysis of the *home advantage* from spatio-temporal data

(iv) Using the aligned roles, a technique was proposed to de-noise noisy tracking data using a bilinear spatio-temporal basis model. The spatial basis of the signal were represented using the aligned roles, and discrete cosine transform (DCT) coefficients were used for the temporal component. From this, the underlying signal of group movements was modelled to provide a compact signal to perform clustering and analysis of group behaviours even in the presence of noise. This enabled common formations and spatio-temporal patterns of groups to be discovered.

(v) A real-time system to recognise group activities using macroscopic approaches of centroids and occupancy maps to represent and align the multi-agent data, and were shown to be able to detect group activities effectively even in the presence of noise.

(vi) A new database for evaluating person re-identification models in real-life conditions was presented together with an evaluation protocol to evaluate what factors affect feature performance.

(vii) The use of group information in the form of roles was proposed to improve person re-identification. This was found to be effective especially in low-resolution video footage where people wear similar clothing, and traditional appearance-based approaches would fail.

8.2 Future Work

In this thesis novel methods for aligning group behaviours were proposed to enable large-scale analysis of group behaviours. While these enabled a lot of interesting analysis to be performed, the proposed approaches could be extended to further improve performance. In particular, simple classification, clustering and fusion methods were used to demonstrate the utility of the proposed approaches. These were successful in highlighting the analysis made possible with the proposed representations, but more complex methods could be used to further boost performance. Deep learning methods have grown in popularity in recent years and currently achieve state-of-the-art performance in many domains including speech recognition [54] and visual object recognition [78]. It would be interesting to see how given more labelled training data, deep learning could be applied to discover the inherent features of groups, and to see how well these correspond to the features proposed in this work. In addition, reinforcement learning approaches which relate to how agents ought to take actions in an environment so as to maximise some notion of a cumulative reward, could be explored.

Delving deeper into the various strategic patterns that groups exhibit is another direction for future research. For example, the proposed alignment approaches and formation representation could be considered in sports for short-term prediction (e.g. who will the next pass go to), as well as longer-term prediction (e.g. the match result). The proposed techniques could also be further extended and incorporated into a system for real-time in-game analysis.

Bibliography

[1] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Nonrigid structure from motion in trajectory space," in *Advances in Neural Information Processing Systems (NIPS)*, 2008. 97

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Trajectory space: A dual representation for nonrigid structure from motion," *Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2010. 98

[3] I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh, "Bilinear spatiotemporal basis models," *ACM Transactions on Graphics (TOG)*, 2012. 8, 86, 96, 98, 104

[4] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *European Conference on Computer Vision (ECCV)*, 2008. 20, 21

[5] R. Almeida, L. Reis, and A. Jorge, "Analysis and forecast of team formation in the simulated robotic soccer domain," in *Progress in Artificial Intelligence*. Springer, 2009. 24

[6] L. O. Alvares, V. Bogorny, B. Kuijpers, J. de Macedo, B. Moelans, and A. Vaisman, "A model for enriching trajectories with semantic geographical information," in *Advances in Geographic Information Systems (GIS)*, 2007. 17

[7] O. Arikan, "Compression of motion capture databases," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, 2006. 98

[8] D. Ashbrook and T. Starner, "Using GPS to learn significant locations and predict movement across multiple users," *Personal and Ubiquitous Computing*, vol. 7, no. 5, 2003. 16

[9] I. Atmosukarto, B. Ghanem, S. Ahuja, K. Muthuswamy, and N. Ahuja, "Automatic recognition of offensive team formation in American Football plays," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013. 23

[10] H. Ayanegui and F. Ramos, "Recognizing patterns of dynamic behaviors based on multiple relations in soccer robotics domain," in *Pattern Recognition and Machine Intelligence*. Springer, 2007. 24

[11] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using Haar-based and DCD-based signature," in *Advanced Video and Signal-Based Surveillance (AVSS)*, 2010. 126, 130

[12] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Multiple-shot human re-identification by mean riemannian covariance grid," in *Advanced Video and Signal-Based Surveillance (AVSS)*, 2011. 127, 151

[13] B. Banerjee, L. Kraemer, and J. Lyle, "Multi-agent plan recognition: Formalization and algorithms," in *AAAI Conference on Artificial Intelligence*, 2010. 110

[14] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *European Conference on Computer Vision (ECCV)*, 2006. 126

[15] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by HPE signature," in *International Conference on Pattern Recognition (ICPR)*, 2010. 127

[16] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, "The R*-tree: an efficient and robust access method for points and rectangles," in *International Conference on Management of Data (ICMD)*, no. 2, 1990. 19

[17] M. Beetz, N. von Hoyningen-Huene, B. Kirchlechner, S. Gedikli, F. Siles, M. Durus, and M. Lames, "ASPOGAMO: Automated sports game analysis models," *International Journal of Computer Science in Sport*, vol. 8, no. 1, 2009. 25, 111

[18] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 4, 2002. 40

[19] S. Berchtold, D. Keim, and H.-P. Kriegel, "The X-tree: An index structure for high-dimensional data," *Readings in multimedia computing and networking*, vol. 451, 2001. 19

[20] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," *Data and Knowledge Engineering*, vol. 60, no. 1, 2007. 17, 19

[21] H. Bouma, S. Borsboom, R. J. M. den Hollander, S. H. Landsmeer, and M. Worring, "Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination," in *SPIE Defense, Security, and Sensing*, vol. 8359, 2012. 130

[22] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Computer Vision and Pattern Recognition (CVPR)*, 2000. 97

[23] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *International Conference on Computer Vision (ICCV)*, 2009. 20

[24] J.-C. Bricola, "Classification of multi-agent trajectories," Master's thesis, Ecole polytechnique federale de Lausanne (EPFL), 2012. 27

[25] A. Bronstein, M. Bronstein, and R. Kimmel, *Numerical geometry of non-rigid shapes*. Springer, 2008. 96

[26] Y. Cai, N. de Freitas, and J. Little, "Robust visual tracking for multiple targets," in *European Conference on Computer Vision (ECCV)*. Springer, 2006. 20

[27] P. Carr, M. Mistry, and I. Matthews, "Hybrid robotic/virtual pan-tilt-zoom cameras for autonomous event recording," in *ACM Multimedia*, 2013. 22, 27

[28] P. Carr, Y. Sheikh, and I. Matthews, "Monocular object detection using 3D geometric primitives," in *European Conference on Computer Vision (ECCV)*, 2012. 88, 90, 148

[29] D. Cervone, A. D'Amour, L. Bornn, and K. Goldsberry, "POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data," in *MIT Sloan Sports Analytics Conference*, 2014. 26

[30] M. Chang, N. Krahnstoever, and W. Ge, "Probabilistic group-level motion analysis and scenario recognition," in *International Conference on Computer Vision (ICCV)*, 2011. 110

[31] Z. Chen, H. Shen, and X. Zhou, "Discovering popular routes from trajectories," in *International Conference on Data Engineering (ICDE)*. IEEE, 2011. 17

[32] A. Cheriyadat and R. Radke, "Automatically determining dominant motions in crowded scenes by clustering partial feature trajectories," in *International Conference on Distributed Smart Cameras (ICDSC)*, 2007. 22

[33] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and applications," *Computer Vision and Image Understanding (CVIU)*, vol. 61, no. 1, 1995. 97

[34] M. Cox, S. Sridharan, S. Lucey, and J. Cohn, "Least squares congealing for unsupervised alignment of images," in *Computer Vision and Pattern Recognition (CVPR)*, 2008. 28

[35] A. Criminisi, I. Reid, and A. Zisserman, "A plane measuring device," *Image and Vision Computing*, vol. 17, no. 8, 1999. 90

[36] Q. Dai, P. Carr, L. Sigal, and D. Hoiem, "Family member identification from photo collections," in *Winter Conference on Applications of Computer Vision (WACV)*, 2015. 153

[37] F. De la Torre and M. Black, "Robust parameterized component analysis," in *European Conference on Computer Vision (ECCV)*. Springer, 2002. 28

[38] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, 1977. 59

[39] S. Denman, A. Bialkowski, C. Fookes, and S. Sridharan, "Determining operational measures from multi-camera surveillance systems using soft biometrics," in *Advanced Video and Signal-Based Surveillance (AVSS)*, 2011. 135

[40] S. Denman, A. Bialkowski, C. Fookes, and S. Sridharan, "Identifying customer behaviour and dwell time using soft biometrics," *Video Analytics for Business Intelligence (VABI)*, 2012. 136

[41] S. Denman, C. Fookes, A. Bialkowski, and S. Sridharan, "Soft-biometrics: unconstrained authentication in a surveillance environment," in *Digital Image Computing: Techniques and Applications (DICTA)*, 2009. 126, 130

[42] S. Denman, C. Fookes, and S. Sridharan, "Improved simultaneous computation of motion detection and optical flow for object tracking," in *Digital Image Computing: Techniques and Applications (DICTA)*, 2009. 134

[43] T. D'Orazio and M. Leo, "A review of vision-based systems for soccer video analysis," *Pattern Recognition*, vol. 43, no. 8, 2010. 25, 111

[44] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *International Conference on Computer Vision (ICCV)*, 2007. 130

[45] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR)*, 2010. 127, 130, 149

[46] J. Ferryman, Ed., *Proc. of PETS2006*, 2006. 130

[47] P. Forssén, "Maximally stable colour regions for recognition and matching," in *Computer Vision and Pattern Recognition (CVPR)*, 2007. 127

[48] W. Förstner and B. Moonen, "A metric for covariance matrices," *Technical Report, University of Stuttgart*, 1999. 150

[49] B. Frey and N. Jojic, "Transformed component analysis: Joint estimation of spatial transformations and image components," in *International Conference on Computer Vision (ICCV)*, vol. 2, 1999. 28

[50] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Computer Vision and Pattern Recognition (CVPR)*, 2006. 126, 130

[51] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti, "Unveiling the complexity of human mobility by querying

and mining massive trajectory data," *The International Journal on Very Large Data Bases (VLDB)*, vol. 20, no. 5, 2011. 17

[52] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Knowledge discovery and data mining (KDD)*, 2007. 17

[53] K. Goldsberry, "CourtVision: New visual and spatial analytics for the NBA," in *MIT Sloan Sports Analytics Conference*, 2012. 26

[54] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013. 164

[55] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *European Conference on Computer Vision (ECCV)*, 2008. 126, 130, 139, 157

[56] J. Gudmundsson and M. Krevald, "Computing longest duration flocks in trajectory data," in *Advances in Geographic Information Systems (GIS)*, 2006. 16

[57] J. Gudmundsson and T. Wolle, "Football analysis using spatio-temporal tools," *Computers, Environment and Urban Systems*, 2013. 26

[58] A. Gupta, P. Srinivasan, J. Shi, and L. Davis, "Understanding videos, constructing plots: Learning a visually grounded storyline model from annotated videos," in *Computer Vision and Pattern Recognition (CVPR)*, 2009. 25, 111

[59] R. Güting, T. Behr, and J. Xu, "Efficient k-nearest neighbor search on moving object trajectories," *The International Journal on Very Large Data Bases (VLDB)*, vol. 19, no. 5, 2010. 19

[60] A. Guttman, "R-trees: a dynamic index structure for spatial searching," in *International Conference on Management of Data (ICMD)*, no. 2, 1984. 19

[61] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," in *International Conference on Distributed Smart Cameras (ICDSC)*, 2008. 126

[62] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, 1995. 21

[63] A. Hervieu and P. Bouthemy, "Understanding sports video using players trajectories," in *Intelligent Video Event Analysis and Understanding*, J. Zhang, L. Shao, L. Zhang, and G. Jones, Eds. Springer Berlin / Heidelberg, 2010. 111

[64] R. Hess and A. Fern, "Discriminatively trained particle filters for complex multi-object tracking," in *Computer Vision and Pattern Recognition (CVPR)*, 2009. 111

[65] R. Hess, A. Fern, and E. Mortensen, "Mixture-of-parts pictorial structures for objects with variable part sets," in *International Conference on Computer Vision (ICCV)*, 2007. 24, 111

[66] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," *Image Analysis*, 2011. 130

[67] M. Hu, W. Hu, and T. Tan, "Tracking people through occlusions," in *International Conference on Pattern Recognition (ICPR)*, 2004. 126

[68] C. Huang, H. Shih, and C. Chao, "Semantic analysis of soccer video using dynamic bayesian networks," *T. Multimedia*, vol. 8, no. 4, 2006. 25, 111

[69] S. Intille and A. Bobick, "A framework for recognizing multi-agent action from visual evidence," in *AAAI Conference on Artificial Intelligence*, 1999. 110

[70] S. Intille and A. Bobick, "Recognizing planned, multi-person action," *Computer Vision and Image Understanding (CVIU)*, vol. 81, 2001. 26

[71] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *Biometric Authentication*. Springer, 2004. 125

[72] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *Computer Vision and Image Understanding (CVIU)*, vol. 109, no. 2, Feb. 2008. 128

[73] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *Computer Vision and Pattern Recognition (CVPR)*, 2005. 126

[74] S. Khan and M. Shah, "Detecting group activities using rigidity of formation," in *ACM Multimedia*, 2005. 23

[75] K. Kim, M. Grundmann, A. Shamir, I. Matthews, J. Hodgins, and I. Essa, "Motion fields to predict play evolution in dynamic sports scenes," in *Computer Vision and Pattern Recognition (CVPR)*, 2010. 27, 111

[76] K. Kitani, B. Ziebart, A. Bagnell, and M. Herbert, "Activity forecasting," in *European Conference on Computer Vision (ECCV)*, 2012. 22

[77] F. Klügl and G. Rindsfüser, "Large-scale agent-based pedestrian simulation," in *Multiagent System Technologies*. Springer, 2007. 21

[78] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012. 164

[79] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, 1955. 38, 59, 61, 99, 154

[80] M. Lazarescu and S. Venkatesh, "Using camera motion to identify different types of American Football plays," in *International Conference on Multimedia and Expo (ICME)*, 2003. 25, 111

[81] E. G. Learned-Miller, "Data driven image models through continuous joint alignment," *Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 28, no. 2, 2006. 28

[82] J. Lee, J. Han, and K. Whang, "Trajectory clustering: A partition-and-group framework," in *International Conference on Management of Data (ICMD)*, 2007. 16, 17

[83] Y. Lee and S. Choi, "Minimum entropy, k-means, spectral clustering," in *International Joint Conference on Neural Networks*, 2004. 8, 54, 58

[84] B. Leibe, K. Schindler, N. Cornelius, and L. van Gool, "Coupled detection and tracking from static cameras and moving vehicles," *Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, no. 10, 2008. 20

[85] R. Li and R. Chellappa, "Group motion segmentation using a spatio-temporal driving force model," in *Computer Vision and Pattern Recognition (CVPR)*, 2010. 111

[86] R. Li, R. Chellappa, and S. Zhou, "Learning multi-modal densities on discriminative temporal interaction manifold for group activity recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2009. 110

[87] G. Lian, J. Lai, and W. S. Zheng, "Spatial–temporal consistent labeling of tracked pedestrians across non-overlapping camera views," *Pattern Recognition Letters*, vol. 44, no. 5, 2011. 128

[88] D. Lin, E. Grimson, and J. Fisher, "Learning visual flows: A Lie algebraic approach," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009. 20

[89] C. Liu, S. Gong, C. Loy, and X. Lin, "Person re-identification: What features are important?" in *European Conference on Computer Vision (ECCV) Workshops and Demonstrations*, 2012. 127

[90] J. Liu, P. Carr, R. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *Computer Vision and Pattern Recognition (CVPR)*, 2013. 128

[91] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Computer Vision and Pattern Recognition (CVPR)*, 2009. 23

[92] T. Liu, W. Ma, and H. Zhang, "Effective feature extraction for play detection in American Football video," in *Multimedia Modelling Conference (MMM)*, 2005. 25, 111

[93] W. Lu, J. A. Ting, K. P. Murphy, and J. J. Little, "Identifying players in broadcast sports videos using conditional random fields," in *Computer Vision and Pattern Recognition (CVPR)*, 2011. 129

[94] P. Lucey, A. Bialkowski, P. Carr, E. Foote, and I. Matthews, "Characterizing multi-agent team behavior from partial team tracings: Evidence from the English Premier League," in *AAAI Conference on Artificial Intelligence*, 2012. 26, 70, 72

[95] P. Lucey, A. Bialkowski, P. Carr, S. Morgan, I. Matthews, and Y. Sheikh, "Representing and discovering adversarial team behaviors using player

roles," in *Computer Vision and Pattern Recognition (CVPR)*, 2013. 39, 129, 153

[96] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews, "Assessing team strategy using spatiotemporal data," in *Knowledge discovery and data mining (KDD)*, 2013. 26, 70, 72, 78

[97] C. Madden, M. Piccardi, and S. Zuffi, "Comparison of techniques for mitigating the effects of illumination variations on the appearance of human targets," in *Advances in Visual Computing*. Springer, 2007. 126

[98] R. Masheswaran, Y. Chang, J. Su, S. Kwok, T. Levy, A. Wexler, and N. Hollingsworth, "The three dimensions of rebounding," in *MIT Sloan Sports Analytics Conference*, 2014. 26

[99] R. Mazzon, S. F. Tahir, and A. Cavallaro, "Person re-identification in crowd," *Pattern Recognition Letters*, vol. 33, no. 14, 2012. 128

[100] A. Miller, L. Bornn, R. Adams, and K. Goldsberry, "Factorized point process intensities: A spatial analysis of professional basketball," in *International Conference on Machine Learning (ICML)*, 2014. 26

[101] A. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, 2008. 25, 111

[102] V. Morariu and L. Davis, "Multi-agent event recognition in structured scenarios," in *Computer Vision and Pattern Recognition (CVPR)*, 2011. 111

[103] B. Morris and M. Trivedi, "Learning trajectory patterns by clustering: Experimental studies and comparative evaluation," in *Computer Vision and Pattern Recognition (CVPR)*, 2009. 18

[104] T. Nakashima, T. Uenishi, and Y. Narimoto, "Off-line learning of soccer formations from game logs," in *World Automation Congress (WAC)*. IEEE, 2010. 24

[105] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2002. 137

[106] A. Palma, V. Bogorny, B. Kuijpers, and L. Alvares, "A clustering-based approach for discovering interesting places in trajectories," in *ACM Symposium on Applied Computing*, 2008. 17

[107] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR)*, 2009. 20, 21

[108] J. Peña and H. Touchette, "A network theory analysis of football strategies," *arXiv preprint arXiv:1206.6904*, 2012. 26

[109] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 34, no. 11, 2012. 28

[110] M. Perse, M. Kristan, S. Kovacic, and J. Pers, "A trajectory-based analysis of coordinated team activity in basketball game," *Computer Vision and Image Understanding (CVIU)*, 2008. 26, 111

[111] A. Pozo, J. Gracía, M. A. Patricio, and J. M. Molina, "A structured representation to the group behavior recognition issue," in *User-Centric Technologies and Applications*. Springer, 2011. 24

[112] B. Prosser, W. S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *British Machine Vision Conference (BMVC)*, 2010. 127, 130

[113] Prozone, www.prozonesports.com. 33, 43, 55

[114] Z. Qin and C. Shelton, "Improving multi-target tracking via social grouping," in *Computer Vision and Pattern Recognition (CVPR)*, 2012. 22

[115] F. Ramos and H. Ayanegui, "Tracking behaviours of cooperative robots within multi-agent domains," in *Autonomous Agents*, 2010. 24

[116] K. Rao and P. Yip, *Discrete cosine transform: Algorithms, advantages, applications*. New York, NY: Academic, 1990. 98

[117] L. Reis, R. Lopes, L. Mota, and N. Lau, "Playmaker: Graphical definition of formations and setplays," in *Information Systems and Technologies (CISTI)*. IEEE, 2010. 24

[118] S. Roberts, R. Everson, and I. Rezek, "Minimum entropy data partitioning," *International Conference on Artificial Neural Networks (ICANN)*, 1999. 54, 58

[119] M. Rodriguez, I. Laptev, J. Sivic, and J. Audibert, "Density-aware person detection and tracking in crowds," in *International Conference on Computer Vision (ICCV)*, 2011. 20

[120] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert, "Data-driven crowd analysis in video," in *International Conference on Computer Vision (ICCV)*, 2011. 20

[121] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal on Computer Vision (IJCV)*, 2000. 48, 64

[122] A. Sadilek and H. Kautz, "Recognizing multi-agent activities from GPS data," in *AAAI Conference on Artificial Intelligence*, 2008. 110

[123] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis, "Human detection using partial least squares analysis," in *International Conference on Computer Vision (ICCV)*, 2009. 127

[124] B. Siddiquie, Y. Yacoob, and L. Davis, "Recognizing plays in American Football videos," University of Maryland, Tech. Rep., 2009. 111

[125] STATS SportsVU, www.sportvu.com. 33

[126] D. Stracuzzi, A. Fern, K. Ali, R. Hess, J. Pinto, N. Li, T. Konik, and D. Shapiro, "An application of transfer to American Football: From observation of raw video to control in a simulated environment," *AI Magazine*, vol. 32, no. 2, 2011. 26, 111

[127] G. Sukthankar and K. Sycara, "Hypothesis pruning and ranking for large plan recognition problems," in *AAAI Conference on Artificial Intelligence*, 2008. 110

[128] G. Sukthankar and K. Sycara, "Activity recognition for dynamic multi-agent teams," *ACM Trans. Intelligent Systems Technology*, 2012. 110

[129] F. Tang, S. Lim, and N. Chang, "An improved local feature descriptor via soft binning," in *International Conference on Image Processing (ICIP)*, 2010. 126

[130] L. Tang, X. Yu, S. Kim, J. Han, C. Hung, and W. Peng, "Tru-Alarm: Trustworthiness analysis of sensor networks in cyber-physical systems," in *International Conference on Data Mining (ICDM)*, 2010. 16

[131] L. Tang, Y. Zheng, X. Xie, J. Yuan, X. Yu, and J. Han, "Retrieving k-nearest neighbor trajectories by a set of point locations," in *International*

Symposium on Advances in Spatial and Temporal Databases (SSTD), 2011. 16, 19

[132] K. Teknomo, "Microscopic pedestrian flow characteristics: Development of an image processing data collection and simulation model," Ph.D. dissertation, Tohoku University, 2002. 20

[133] A. Torralba, "Contextual priming for object detection," *International Journal of Computer Vision (IJCV)*, vol. 53, no. 2, 2003. 22

[134] L. Torresani and C. Bregler, "Space-time tracking," in *Computer Vision and Pattern Recognition (CVPR)*, 2002. 97

[135] R. Y. Tsai, "An efficient and accurate camera calibration technique for 3D machine vision," in *Computer Vision and Pattern Recognition (CVPR)*, 1986. 132

[136] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *European Conference on Computer Vision (ECCV)*, 2006. 127, 150

[137] UK Home Office, "Imagery library for intelligent detection systems (i-LIDS): Multiple camera tracking scenario definition," 2008. [Online]. Available: www.homeoffice.gov.uk/science-research/hosdb/i-lids/ 130

[138] X. Wei, P. Lucey, S. Morgan, and S. Sridharan, "Predicting shot locations in tennis using spatiotemporal data," in *Digital Image Computing: Techniques and Applications (DICTA)*, 2013. 26

[139] X. Wei, P. Lucey, S. Morgan, and S. Sridharan, "Sweet-spot: Using spatiotemporal data to discover and predict shots in tennis," in *MIT Sloan Sports Analytics Conference*, 2013. 26

[140] G. Wu, A. Rahimi, E. Chang, K. Goh, T. Tsai, A. Jain, and Y. Wang, "Identifying color in motion in video sensors," in *Computer Vision and Pattern Recognition (CVPR)*, 2006. 126, 135

[141] C. Xu, Y. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using webcast text for semantic event detection in broadcast," *T. Multimedia*, vol. 10, no. 7, 2008. 25, 111

[142] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *International Conference on Pattern Recognition (ICPR)*, 2006. 130

[143] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-Drive: Driving directions based on taxi trajectories," in *Advances in Geographic Information Systems (GIS)*, 2010. 16

[144] L. Zhang, Y. Li, and R. Nevatia, "Global data associated," in *Computer Vision and Pattern Recognition (CVPR)*, 2008. 20

[145] Y. Zhang, W. Ge, M. Chang, and X. Liu, "Group context learning for event recognition," in *Workshop on Applications of Computer Vision (WACV)*, 2012. 22, 110

[146] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2013. 127

[147] W. S. Zheng, S. Gong, and T. Xiang, "Quantifying contextual information for object detection," in *International Conference on Computer Vision (ICCV)*, 2009. 22

[148] W. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *British Machine Vision Conference (BMVC)*, vol. 5, London, UK, 2009. 22, 128, 129, 145

[149] Y. Zheng, X. Xie, and W. Ma, "GeoLife: A collaborative social networking service among user, service," *IEEE Data Engineering Bulletin*, 2010. 16

[150] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Computer Vision and Pattern Recognition (CVPR)*, 2012. 21

[151] C. Zhou, N. Bhatnagar, S. Shekhar, and L. Terveen, "Mining personally important places from GPS tracks," in *International Conference on Data Engineering (ICDE) Workshop*, 2007. 16

[152] G. Zhu, Q. Huang, C. Xu, Y. Rui, S. Jiang, W. Gao, and H. Yao, "Trajectory based event tactics analysis in broadcast sports video," in *ACM Multimedia*, 2007. 26